ENTWICKLUNG VON METHODEN FÜR DAS COMPUTERGESTÜTZTE DESIGN VON MIMOTOPEN

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

Dr. rer. nat.

der Fakultät für
Biologie
an der

Universität Duisburg-Essen

vorgelegt von

Dipl.-Bioinf. (FH) Stanislav Jakuschev

geb. in Taschkent, Usbekistan

Juli 2015

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden am Zentrum für Medizinische Biotechnologie (ZMB) in der Abteilung für Bioinformatik der Universität Duisburg-Essen durchgeführt.

1. Gutachter: Herr Prof. Dr. Daniel Hoffmann

2. Gutachter: Herr Prof. Dr. Peter Bayer

Vorsitzender des Prüfungsausschusses: Herr Prof. Dr. Ralf Küppers

Tag der mündlichen Prüfung: 6. Juli 2015

Entwicklung von Methoden für das computergestützte Design von Mimotopen, EpitopeMatch ©, Stanislav Jakuschev, Juli 2015

Ich widme diese Arbeit: meinen wunderbaren Kindern Kiana und Roland; meiner wunderschönen Frau Irina; meinen fleißigen und kreativen Eltern Rita und Eugen; meinen Cousins Philipp † und Paul †, die für immer in meinem Herzen bleiben; meiner Tante Irina und meinem Onkel Oleg, die immer für mich da sind.

Möge diese Arbeit ihren Beitrag im Feld der Bioinformatik leisten.

DANKSAGUNG

Ich bedanke mich hiermit ganz herzlich bei allen, die meine Promotion ermöglicht und unterstützt haben.

Mein größter Dank gehört meinem Lehrer und Vorbild Prof. Dr. Daniel Hoffmann, für sein Vertrauen in meine Fähigkeiten, für seine niemals endende Geduld und Zuversicht, für seine immer konstruktiven und kostbaren Ratschläge.

Ich danke der Universität Duisburg-Essen, und insbesondere dem Zentrum für Medizinische Biotechnologie, für die stets schöne Arbeitsatmosphäre auf dem höchsten Niveau.

Ein großer Dank für die gute Zeit gilt meinen lieben Kolleginnen und Kollegen Rocio Rebollido-Rios, Stefan Franke, Manuel Prinz, Bjoern Thorwith, Oliver Kuhn, Karsten Sewczyk, J. Nikolaj Dybowski, Dominik Heider und Claudia Wilmes.

Im speziellen bei Rocio Rebollido-Rios bedanke ich mich für ihre Courage beim Einsetzen von EpitopeMatch.

Felix Hess spreche ich meinen herzlichen Dank aus, für seinen kritischen Blick auf das Online-Tutorial von EpitopeMatch.

ZUSAMMENFASSUNG

Die wachsende Menge an experimentell aufgeklärten Protein-Protein-Komplexen oder allgemeiner: Protein-Ligand-Komplexen, erlaubt das immer genauere Studium von biomolekularen Wechselwirkungen. Eine Teilmenge der existierenden Wechselwirkungen bilden die für die adaptive Immunantworten wichtigen Antigen-Antikörper-Wechselwirkungen. Die chemischen Gruppen an der Oberfläche der Antigene entscheiden über die spezifischen Wechselwirkungen mit Antikörpern, und werden als antigene Determinanten oder Epitope bezeichnet. Die dazu komplementären Bindestellen auf den Antikörpern werden als Paratope bezeichnet. Häufig verwendet man den Begriff „Epitop“ allgemein für Molekülteile, die spezifisch erkannt werden. Die Spezifität der Epitope wird sowohl durch die geometrische Anordnung als auch durch die chemische Konfiguration der monomeren Gruppen bestimmt. Mimotope sind synthetisch hergestellte Proteine, die die strukturellen Erkennungsmerkmale der Epitope nachahmen und somit z.B. eine definierte Immunantwort auslösen können.

Beispielsweise ist es nun möglich, Epitope bis zur atomaren Auflösung zu identifizieren und nach ähnlichen Strukturmotiven auf anderen Proteinstrukturen zu suchen. Diese Art des Strukturvergleichs eröffnet interessante Anwendungen: Epitope lassen sich ggf. auf andere Trägermoleküle transplantieren, oder es könnten Kreuzreaktivitäten vorhergesagt werden. Entscheidend für diese Ansätze ist die Verfügbarkeit einer Methode, mit der sich Strukturmotive schnell und genau vergleichen lassen. Die Entwicklung einer solchen Methode (EpitopeMatch) ist das Ziel dieser Promotionsarbeit. Im Einzelnen soll EpitopeMatch folgende Eigenschaften besitzen:

- Einbeziehung geometrischer und chemischer Ähnlichkeit.
- Flexible Definition von i. Allg. diskontinuierlichen Epitopen auf der Grundlage bekannter Komplexstrukturen.
- Effiziente Suche auf großen Strukturdatenbanken.
- Möglichkeit der Transplantation vollständiger Epitope.
- Verknüpfung der Fundstellen mit funktionellen biologischen Daten.

PUBLIKATIONEN

[1] Thilo Albert, Christian Egler, **Stanislav Jakuschev**, Ulrike Schuldenzucker, Alexandra Schmitt, Oliver Brokemper, Martin Zabe-Kühn, Daniel Hoffmann, Johannes Oldenburg, and Rainer Schwaab. The b-cell epitope of the monoclonal anti-factor viii antibody esh8 characterized by peptide array analysis. *Thromb Haemos*t, 99(3):634–637, Mar 2008. doi: 10.1160/TH07-06-0400. URL http://dx.doi.org/10.1160/TH07-06-0400.

[2] **Stanislav Jakuschev** and Daniel Hoffmann. A novel algorithm for macromolecular epitope matching. *algorithms*, 2:498–517, Mar 2009. URL http://www.mdpi.com/1999-4893/ 2/1/498.

[3] Stefanie Ohlig, Pershang Farshi, Ute Pickhinke, Johannes van den Boom, Susanne Höing, **Stanislav Jakuschev**, Daniel Hoffmann, Rita Dreier, Hans R. Schöler, Tabea Dierker, Christian Bordych, and Kay Grobe. Sonic hedgehog shedding results in functional activation of the solubilized protein. *Dev Cell*, 20(6):764–774, Jun 2011. doi: 10.1016/j.devcel. 2011.05.010. URL http://dx.doi.org/10.1016/j.devcel.2011.05.010.

[4] Rocio Rebollido-Rios, Shyam Bandari, Christoph Wilms, **Stanislav Jakuschev**, Andrea Vortkamp, Kay Grobe, and Daniel Hoffmann. Signaling domain of sonic hedgehog as cannibalistic calcium-regulated zinc-peptidase. *PLoS Comput Biol*, 10(7):e1003707, Jul 2014. doi: 10.1371/journal.pcbi.1003707. URL http://dx.doi.org/10.1371/journal.pcbi.1003707.

Poster

[1] J N Dybowski, S Franke, D Heider, D Hoffmann, **S Jakuschev**, O Kuhn, M Prinz, R Rawi, R Rebollido-Rios, K Sewchyk, B Thorwirth, C Wilms, J Winkler. University of Duisburg-Essen, Essen, Germany. Computational Biomolecular Modelling for Functional Analysis and Protein Engineering Bioinformatics.

[2] Rebollido-Rios R, **Jakuschev S**, Hoffmann D. 2012. RECOMB 2012. Barcelona, Spain. How does the calcium binding could affect the suspected protease activity of Sonic Hedgehog?

[3] Rebollido-Rios R, Wilms C, **Jakuschev S**, Dybowski J N, Hoffmann D. 2013. ISMB/ECCB. Berlin, Germany. Molecular modeling of Sonic Hedgehog binding suggests novel regulatory mechanism.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

AC hydrogen-bond ACceptor

AFP Aligned Fragment Pair

AL hydrophobic ALiphatic

APBS Adaptive Poisson-Boltzmann Software

ATOMS atoms

AVEHYDROP AVErage HYDROPathy

BB Branch and Bound

BCS Best Common Substructure

BCSS Best Common Substructure Score

BCSST Best Common Substructure Score Threshold

BCST Best Common Substructure Threshold

BFTS Breadth-First Tree Search

BP Biological Process

BSDB Binding Sites Data Base

CATH Class Architecture Topology Homologous superfamily

CCD Chemical Component Dictionary

CD Clique Detection

CE Combinatorial Extension

CF Contact Face

CP Circular Permutation

COMPL completeness

COMPS components

CS Common Substructure

CSS Common Substructure Score

DFLX Domain FLeXibility

DMCSS Domain Maximum Common Substructure Score

DA mixed Donor/Acceptor

DB Disulfide Bridge

DO hydrogen-bond DOnor

DRGD Domain RiGiDity

DRMSD Domain Root Mean Square Deviation

DS Domain Score

DP Dynamic Programming

EDB Epitope Data Base

EP Electrostatic Potential

FE Feature Extraction

FLX FLeXibility

FN False Negative

FP False Positive

FSSP Family of Structurally Similar Proteins

GA Genetic Algorithm

GH Geometric Hashing

GO Gene Ontology

GSAS Gapped Structural Alignment Score

HCS Homologous Core structure overlap Score

HMMSTR Hidden Markov Model for local STRucture

HS Hinge Score

HSSP Homology-derived StructureS of Proteins

HST Hinge Score Threshold

ID Insertion/Deletion

IDENT identity

LHM Loop Hausdorff Measure

MC Monte Carlo

MCS Maximum Common Substructure

MCSS Maximum Common Substructure Score

MF Molecular Function

MMDB Molecular Modeling DataBase

MOLWEIGHT MOLecular WEIGHT

MOSIEX Molecular Similarity Explorer

MSMS Maximal Speed Molecular Surface

MSTA Multiple STructural Alignment

MTME Markovian Transition Model of Evolution

MVC Model View Control

NBCSNST Normalized Best Common Substructure Node Score Threshold

NBCSS Normalized Best Common Substructure Score

NETCHARGE net charge

NMCSNST Normalized Maximum Common Substructure Node Score Threshold

NMR Nuclear Magnetic Resonance spectroscopy

NP Non-deterministic Polynomial-time

NRPDB Non-Redundant PDB chain set

NRMSD Normalized Root Mean Square Deviation

NWRMSD Normalized Weighted Root Mean Square Deviation

OSF Objective Scoring Function

PAM Point Accepted Mutation

PDB Protein Data Base

PI aromatic PI contact

OF Objective Function

QS Query-Struktur

QMCSS Query Maximum Common Substructure Score

QTMCSS Query-Target Maximum Common Substructure Score

RD Residue Depth

RGD RiGiDity

RIPC Repetitions, Indels, Permutation and Conformational variability

RMSD Root Mean Square Deviation

RMSDT Root Mean Square Deviation Threshold

RMSS Root Mean Square Similarity

RTS Residuum Tuple Score

SA Solvent Accessibility

SAS Solvent Accessible Surface

SCOP Structural Classification Of Proteins

SES Solvent Excluded Surface

SF Spheric Face

ShhN Sonic Hedgehog, N-terminal domain

SM Seed Match

SSE Secondary Structure Element

SSIM Substitution SIMilarty

SCA Structure Comparison and Alignment

TF Toric Face

TN True Negative

TP True Positive

TS Target-Struktur

TMCSS Target Maximum Common Substructure Score

WRMSD Weighted Root Mean Square Deviation

X-ray X-ray diffraction crystallography

Teil I

EINFÜHRUNG

Teil i führt den Leser in die Thematik und die Terminologie der Bioinformatik im Rahmen der Strukturmodellierung und insbesondere des Strukturvergleichs ein. Die in der Form eines kurzen Reviews präsentierten, hierarchisch gegliederten Abschnitte des Kapitel 1 illustrieren den Gedankengang, der zum Design der im Teil ii ausführlich diskutierten Methode EpitopeMatch geführt hat.

1 STRUKTUR

Mimotope sind aus der Sicht der Immunologie entweder die Epitop-nachahmenden, kurzen Peptide oder die Epitop-tragenden, kleinen Proteine [179]. Der Begriff wurde 1986 durch Mario Geysen geprägt und bezog sich damals auf Peptide, die als antigene Determinanten an die Paratope der Antikörper gebunden haben [51]. Ob die Mimotope nun die Epitope oder die Paratope nachbilden, spielt aus der informationstechnischen Sicht keine Rolle. Im Rahmen dieser Arbeit wird jede Wechselwirkungsstelle eines Proteins als ein Epitop und jedes modifizierte Protein als ein Mimotop seines Analogons bezeichnet.

Epitope werden grundsätzlich in zwei Kategorien unterteilt. Die *kontinuierlichen* Epitope sind kurze lineare Peptid-Fragmente, die mit dem Antikörper (oder allg. Wechselwirkungspartner) eine Bindung eingehen können. Wenn die Wechselwirkung mit dem Paratop durch ein einziges Peptid nicht vollständig induziert werden kann, dann muss diese durch weitere Peptide ergänzt werden. Sind mehrere Peptide an der Wechselwirkung mit dem Paratop beteiligt, so bilden sie ein *diskontinuierliches* Epitop nach. Da die einzelnen, voneinander getrennten Peptide über keine native Konformation des Original-Antigens verfügen, weisen sie in der Regel eine schwächere Bindungsaffinität zu dem Antikörper auf, als das Antigen selbst [139]. Je größer das Epitop, desto höher ist die Wahrscheinlichkeit, dass seine Funktion durch die kontinuierliche Konfiguration eines einzigen Peptids nicht ausreichend nachgebildet werden kann, und desto höher ist die Notwendigkeit seine ggf. diskontinuierliche Natur auf eine größere Leitstruktur mit der eigenen nativen Konformation zu übertragen. Trennt man sich von der alleinigen Betrachtung der Antikörper-Antigen-Interaktion und erweitert man sie um die sonstigen Protein-Ligand-Schnittstellen, so wird es schnell deutlich, dass die größte Mehrheit der Epitope bzw. der Wechselwirkungsstellen diskontinuierlich ist.

Im Gegensatz zu der Herstellung der einzelnen Peptide [141] ist die synthetische Rekonstruktion der diskontinuierlichen Epitope deutlich aufwendiger [175]. Diesem gravierenden *in vitro* Defizit wirken die modernen *in silico* Methoden der Bioinformatik entgegen. In dieser Arbeit vorgestelltes, implementiertes und einsatzbereites Verfahren ergänzt die breite Palette der Strukturalignment-Software [178] und erweitert ihre Power um die neuen Ähnlichkeitskoeffizienten.

Die Software EpitopeMatch eröffnet die Möglichkeit der sowohl qualitativen als auch quantitativen Analyse der Ähnlichkeit zwischen den Wechselwirkungsstellen der experimentell aufgeklärten Proteinstrukturen (Protein Data Base (PDB) [24]). Hiermit legt sie das Fundament für die Entwicklung einer neuartigen Datenbank (Abs. 3.2), in der die Proteine nicht nur nach den Ähnlichkeiten der Wechselwirkungsstellen, sondern auch nach ihrem Potential als mögliche Leitstrukturen für die Transplantation der Wechselwirkungsstellen zu dienen, klassifiziert werden können. Solches Netzwerk der Bindungsähnlichkeiten kann z.B. neue Erkenntnisse im Bezug auf die Variabilität der Beschaffenheit der Wechselwirkungsstellen liefern, neue Gemeinsamkeiten zwischen evolutionär weit voneinander entfernten Strukturen aufdecken und einen Beitrag zu einer höheren Auflösung der Beschreibung der jeweiligen Funktion leisten. Neben der mittlerweile auf über 10^5 experimentell aufgeklärten Proteinstrukturen mit über $2.8 \cdot 10^5$ Ketten gewachsenen PDB ist eine schier unendlich große Menge an *in silico* modellierten Strukturen denkbar. Moderne Methoden der evolutionären Paretooptimierung [54, 64, 63] zeigen, dass die Modellierung der kleinen

stabilen Individuen mit transplantierten Epitopen durchaus möglich ist. Die so erzeugten Mimotope sind sowohl aus der medizinischen (bessere Bioverfügbarkeit, kleinere Immunogenität) als auch aus der ökonomischen (geringere Herstellungskosten) Sicht interessant [54]. EpitopeMatch kann im Prozess der Paretooptimierung z.B. als ein Lieferant der Leitstrukturen mit den Vorschlägen für die Transplantationsstellen dienen, die als Startpunkte der evolutionären Paretooptimierung eingesetzt werden können. Während der Erzeugung der neuen, stabilen Individuen kann mittels EpitopeMatch geprüft werden, ob ein Individuum möglicherweise mit einer bekannten Funktion ausgestattet ist, indem das Individuum z.B. gegen die gesamte Datenbank bekannter Epitope gematcht wird. Die Deskriptoren des EpitopeMatch können in diesem Fall für die Messung der Ähnlichkeit zu den bekannten Epitopen eingesetzt werden und würden so als zusätzliche Zielgrößen (objectives) der Paretooptimierung für gerichtete Mutation eingesetzt werden.

1.1 VORHERSAGE

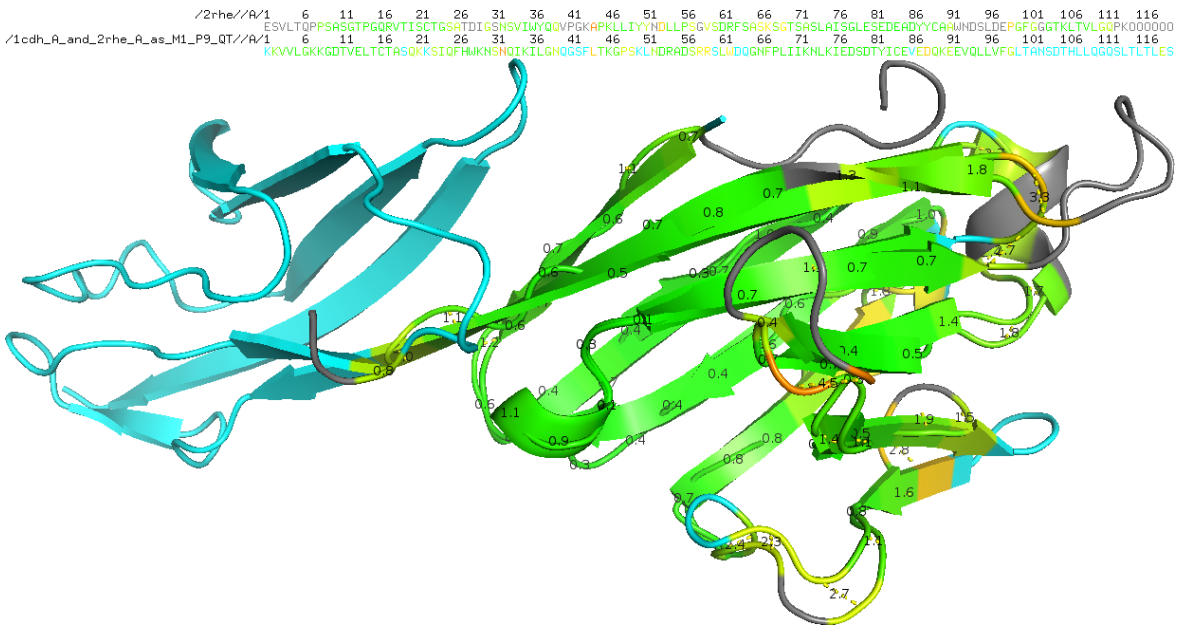
Unter den Makromolekülen zeichnen sich die Proteine durch eine enorme strukturelle und funktionale Diversität aus. Erkenntnisse über sie selbst und ihre Beteiligung an den biologischen Prozessen führen zu dem Verstehen der Funktion bzw. der Fehlfunktion lebender Organismen. Aufgrund der Fortschritte der Sequenzierungstechnologien öffnet sich die Schere zwischen der Anzahl der bekannten Gene und Strukturen stetig weiter. Um diese Lücke zu schließen, gehört die Proteinstrukturvorhersage zu einer der wichtigsten Aufgaben der Strukturbiologie. Darüber hinaus ist das Wissen von der Struktur der Enzyme und Rezeptoren von größter Bedeutung für das Design der Arzneimittel.

1.1.1 *Ab initio Strukturvorhersage*

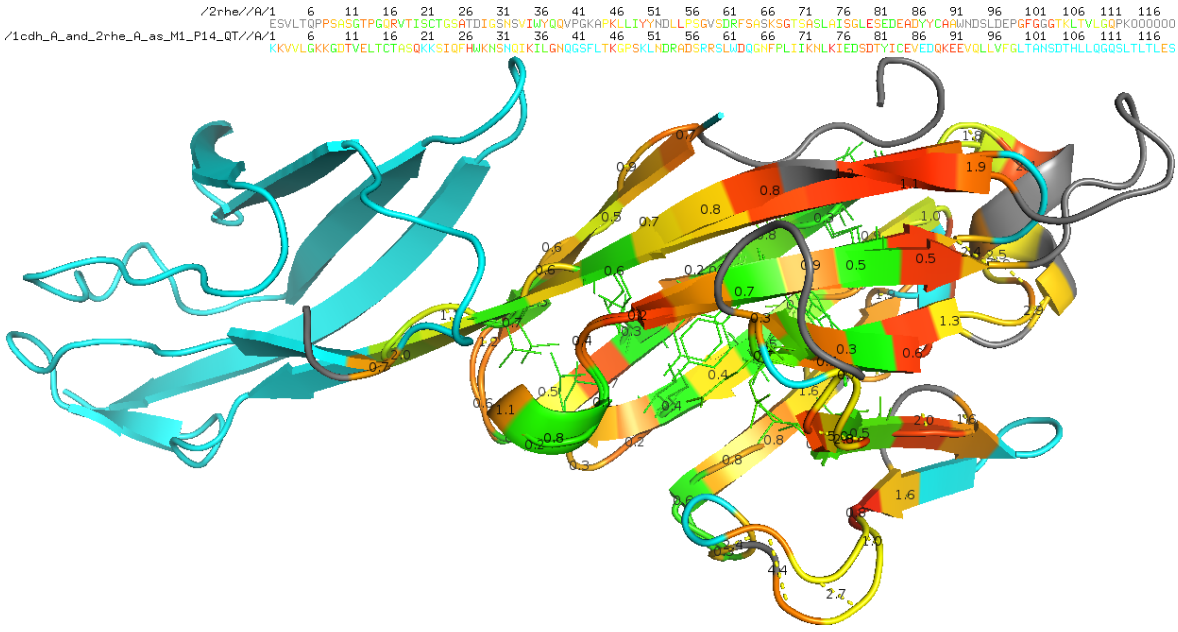
Anfinsen's thermodynamische Hypothese [10] sagt voraus, dass die native Konformation eines Proteins in Lösung in einem unmittelbaren Zusammenhang mit dem globalen Minimum seiner freien Energie steht. Diese fundamentale Erkenntnis führte zu der Entwicklung der sogenannten *ab initio* Methoden. Sie nutzen rechenintensive Strategien [85] für die Nachbildung der an der Proteinfaltung beteiligten physiko-chemischen Phänomene, um die Konformationsräume der Proteine nach den Konformationen mit der niedrigsten freien Energie zu durchsuchen. Die Größe der Konformationsräume aufgrund der dynamischen Natur der Proteine und die Ungenauigkeiten in der Definition der Scoringfunktionen bei der Berücksichtigung der elektrostatischen Effekte und des Lösungsmittels erschweren die Aufgabe beträchtlich [18]. Die aktuellen Fortschritte in der Hardware- und Softwareentwicklung erlauben immer tiefere Einblicke in den noch nicht vollständig aufgeklärten Prozess der Proteinfaltung [132]. Die Rechenleistung konventioneller PCs reicht für die effiziente Berechnung von Strukturen einer typischen Domänengröße der globulären Proteine von ~150 Aminosäuren [186] nicht aus. Obwohl die Mimotope aus immunologischen Gründen so klein wie möglich sein sollten, ist ihre Vorhersage mit den reinen *ab initio* Methoden schwierig und unwirtschaftlich.

1.1.2 *Komparative Strukturvorhersage*

Seit einigen Dekaden leisten die experimentellen Strukturbestimmungsmethoden [185] den wesentlichen Beitrag zur Aufklärung der dreidimensionalen Proteinstruktur. Obwohl die Möglichkeiten dieser Methoden durch die Einschränkungen bezüglich der Kristallisierbar-



(a) Ein Beispiel aus dem Datensatz von 225 repräsentativen Proteinstrukturen [62]. Query-Struktur 1CDH.A (T-Cell Surface Glycoprotein, Kette A, 178 Aminosäuren, cyan) verglichen mit der Target-Struktur 2RHE.A (Immunoglobulin, Kette A, 114 Aminosäuren, grau). EpitopeMatch ermittelte anhand von Rückgratatomen (C, O, Cα, N) 87 Aminosäurenpaare (grün über gelb bis rot / ähnlich bis unähnlich) als die beste gemeinsame Substruktur (BCS) mit einer RMSD von 1.443$\mathring{A}$. Die Fließkommazahlen bezeichnen die Abstände zwischen den Cα-Atomen der korrespondierenden Residuen. Der Prozentsatz der Aminosäurenpaare mit identischen Aminosäuren liegt bei 20.69% (18/87). Die Ähnlichkeit der BCS (reine Geometrie des Rückgrats) beläuft sich auf 92.06%.



(b) Eine Erweiterung der Geometrie des Rückgrats um die physiko-chemischen Eigenschaften der Aminosäuren (Molekulargewicht, Hydrophobizität, Nettoladung) resultiert in einer BCS mit 83 korrespondierenden Aminosäurenpaaren mit einer RMSD von 1.503$\mathring{A}$. Es wird sichtbar, dass der hydrophobe Kern (überwiegend grün, Residuenreste sind als Linien dargestellt) deutlich höher Konserviert ist als die Residuen an der lösungsmittelzugänglichen Oberfläche (überwiegend orange und rot). Der Prozentsatz der Aminosäurenpaare mit identischen Aminosäuren liegt bei diesem Alternativalignment bei 25.3% (21/83). Die Ähnlichkeit der BCS beläuft sich auf 64.67%, deutlich niedriger als im Fall der reinen Geometrie. Wenn man davon ausgeht, dass die Funktion der Proteine unmittelbar mit den Residuen an der lösungsmittelzugänglichen Oberfläche zusammenhängt, dann ist die Divergenz der Funktion in diesem Beispiel deutlich höher als die Divergenz des Faltungsmusters.

Abbildung 1: Sequenzielle und strukturelle Divergenz.

keit (X-ray diffraction crystallography (X-ray)) bzw. der Größe und der Löslichkeit (Nuclear Magnetic Resonance spectroscopy (NMR)) der Proteine begrenzt sind, bildet die Menge der aufgeklärten Proteinstrukturen eine beachtliche Basis für die Strukturvorhersage mittels komparativer Modellierung.

Obwohl die Wahrscheinlichkeit für die Ähnlichkeit zweier Proteinstrukturen mit der steigenden Ähnlichkeit ihrer Sequenzen wächst, oder umgekehrt, die strukturelle Divergenz der homologen Proteine unmittelbar mit den wachsenden Differenzen auf der Sequenzebene zusammenhängt [37], sind weder die unterschiedlichen Faltungsmuster mit ähnlichen Sequenzen noch die ähnlichen Faltungsmuster mit unterschiedlichen Sequenzen ausgeschlossen. Dabei konserviert die Evolution die dreidimensionalen Strukturmuster der Proteine wesentlich höher als die eindimensionalen Muster ihrer Primärstruktur [102]. Strukturen unterhalb der Zwielichtzone der Sequenzidentität von 30% [144] können insgesamt sehr ähnliche Faltungsmuster besitzen (Abb. 1). Trotz der Unschärfe in der Abhängigkeit der Faltungsmuster von den Sequenzen sind die heutigen Methoden der Homologiemodellierung [112, 183] ein probates Mittel der Strukturvorhersage. Der Modellierungsprozess gliedert sich in sogenannte Basisoperationen (unit operations): Sequenz-/Strukturalignment, Strukturmodellierung, Docking, Protein-Engineering [182]. Für jede dieser Basisoperationen existiert eine große Anzahl an Alternativalgorithmen, die sich in ihrer informationstechnischen Umsetzung voneinander unterscheiden. Der Grund für die Existenz der Alternativalgorithmen pro Basisoperation, bzw. für die Nichtexistenz eines alle Basisoperationen umfassenden Algorithmus, ist die Komplexität der Realitätsabbildung. In diesem Kontext ist das EpitopeMatch eine Basisoperation für das Strukturalignment, die als Vor- und Zwischenstufe für die Strukturmodellierung, Docking und Protein-Engineering dienen kann.

1.1.2.1 *Von der Sequenz zu der Struktur*

Ausgehend von der bekannten Sequenz der gesuchten Target-Struktur werden Strukturdatenbanken [24, 69, 128, 71, 9] unter Verwendung der Sequenzalignmentmethoden nach den Template-Strukturen durchsucht. Existiert eine hohe Sequenzähnlichkeit, so kommt man mit dem paarweisen Sequenzalignment [6, 11, 131] am schnellsten ans Ziel. Im Fall einer im Allgemeinen niedrigen Sequenzähnlichkeit kommt das multiple Sequenzalignment [7, 125, 100] zum Einsatz. Die finalen homologen Sequenzen sind dabei das Resultat einer iterativen Erweiterung des initialen Sets um weitere Homologen, bis keine Homologen mehr gefunden werden können. Wenn keine nennenswerten Template-Strukturen mit den Sequenzalignmentmethoden identifiziert werden können, dann wird die Aufgabe von den Threading-Methoden [27, 78, 165] oder Sequenz-Struktur-Alignmentmethoden [105, 69, 164] übernommen. Im ersten Fall wird die Target-Sequenz über die bekannten Faltungsmuster bewegt, wobei die so entstehenden Template-Strukturen nach der strukturabhängigen Funktion bewertet werden. Im zweiten Fall werden die von der Target-Sequenz und den Template-Strukturen abgeleiteten Profile (Sekundärstruktur, Hydrophobizität, etc.) miteinander verglichen.

Ist eine Reihe von Template-Strukturen identifiziert, so werden sie mittels der kontinuierlichen Strukturalignmentmethoden (Abs. 1.2.1) miteinander verglichen. Die daraus resultierenden Gemeinsamkeiten werden als Anhaltspunkte für die Modellbildung verwendet. Die Konstruktion des 3D-Modells der Target-Struktur erfolgt durch die Komposition der konservierten Motive wie Sekundärstrukturen und Domänen [25, 169, 52], der kleineren kontinuierlichen Fragmente [79, 38, 173, 104] oder durch Auswertung und Anwendung der stereochemischen Eigenschaften wie Bindungslängen und -winkel, Diederwinkel und interatomaren Beziehungen [59, 146, 168, 14]. Die Modelle werden somit unter Berücksichtigung der

intrinsischen Eigenschaften der Rückgrat- und Seitenkettenstruktur erzeugt und nachträglich mittels Energieminimierungsverfahren bzw. Molekular-Dynamik-Simulationen [107, 2] verfeinert.

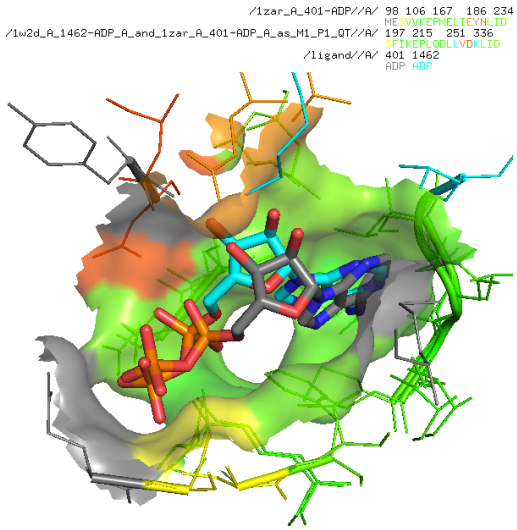
Schnell gelangt man zu der Ansicht, dass erst die Strukturseg mentierung in kleinere kontinuierliche Fragmente die Modellierung von Insertionen, Deletionen und Loops ermöglicht. Die Schwierigkeit der Modellierung erhöht sich drastisch mit den fehlenden Strukturhomologen und mit der steigenden Flexibilität der Fragmente. Insbesondere die Loops [174, 47, 48], die oft für die funktionale Spezifität verantwortlich sind, oder zu der Aktivität der Wechselwirkungsstellen einen Beitrag leisten, unterliegen hohen Mutationsraten und sind als die flexibelsten Fragmente besonders schwer zu modellieren. Die Fehlererkennung übernimmt die Software zur Evaluierung der Proteinstrukturen. Neben der Stereochemie [181, 23] werden auch die Faltungsdichte und Wasserstoffbrückenbindungen [53, 73], Lösungszugänglichkeit und elektrostatisches Potential [30, 92, 44] untersucht. Die erkannten Fehler in Target-Strukturen liefern Hinweise auf die Schwächen im Modellierungsprozess und können z.B. zu der automatischen Neuauswahl besser geeigneter Template-Strukturen führen. Verfeinerung der Target-Strukturen zu den physikalisch realistischen Modellen erfolgt mit auf der atomaren Ebene hochauflösenden Methoden unter Berücksichtigung der Kraftfelder und des Lösungsmittels [96, 97].

1.1.2.2 *Vom Epitop zum Mimotop*

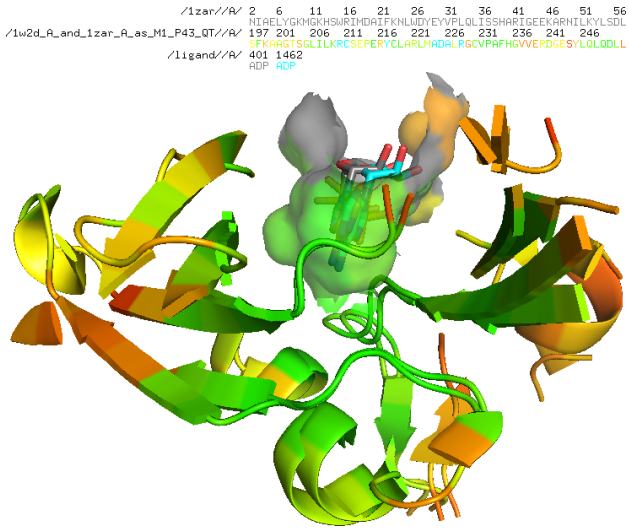
Der Ausgangspunkt ist eine bekannte Struktur eines Epitops, wobei die meisten Epitope diskontinuierlich sind [139] und somit durch eine einzige Sequenz nicht abgebildet werden können. Darüber hinaus steht die Definition einer Wechselwirkungsstelle in einer unmittelbaren Beziehung mit ihrem Interaktionspartner und kann, im Gegensatz zu der Struktur des gesamten Moleküls, nicht ohne der Berücksichtigung des Bindungspartners charakterisiert werden [140]. Ist die Struktur eines Komplexes bekannt, so kann das Epitop mittels der Analyse der interatomaren Protein-Ligand-Kontakte identifiziert und charakterisiert werden [166].

Als Template- bzw. Leitstrukturen können nun alle experimentell bestimmten oder modellierten Strukturen fungieren, die z.B. über eine dem gesuchten Epitop entsprechende räumliche Konfiguration des Proteinrückgrats verfügen, oder aber auch der Seitenketten, unabhängig von der relativen Position der entsprechenden Rückgratatome. Die Suche nach diesen vermutlichen Transplantationsstellen erfolgt mittels der diskontinuierlichen Strukturalignmentmethoden (Abs. 1.2.2). Der Oberflächenanteil einer Wechselwirkungsstelle, der in der Regel die meisten, für den größten Anteil der Bindungsenergie verantwortlichen Hotspot-Residuen präsentiert, und im unmittelbaren Kontakt mit dem Liganden steht, ist für die Nachbildung einer Wechselwirkung nicht ausreichend. Mit der wachsenden Größe einer Wechselwirkungsstelle spielt die Mechanik des tragenden Gerüsts eine immer größer werdende Rolle im Prozess der Assoziation mit dem Liganden und dem damit verbundenen Induced-Fit [95] (Abs. 2.2.7).

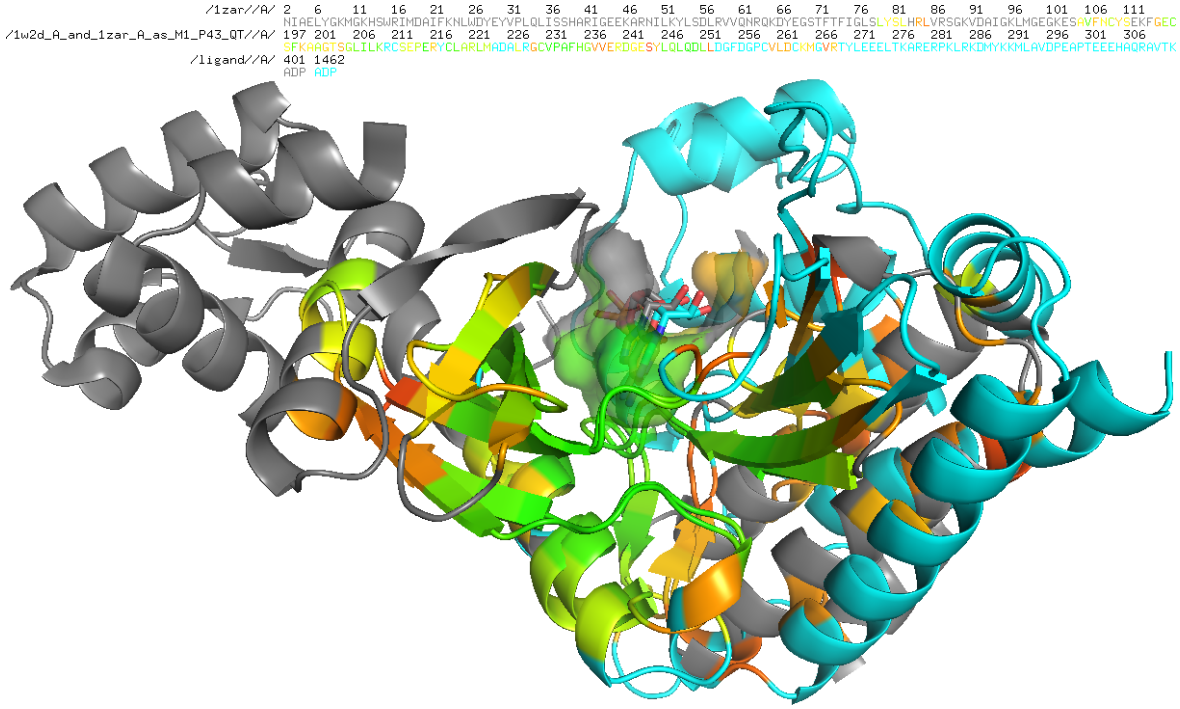
Nach der Transplantation eines Epitops auf eine Leitstruktur muss das so entstandene Rohmimotop, analog zu den oben beschriebenen Target-Strukturen, auf seine Struktur- [53, 73, 181, 23] und Elektrostatik-Eigenschaften [30, 92, 44, 157] überprüft werden, und anschließend verfeinert [96, 97] werden. Da sowohl der Epitop-Ligand-Komplex als auch die Transplantationsstelle auf der Leitstruktur bekannt sind, erlauben die daraus resultierenden Transformationsdaten eine vorläufige Positionierung des Liganden an der Wechselwirkungsstelle des modellierten Mimotops. Ein solcher Mimotop-Ligand-Komplex vereinfacht den Einsatz der Docking-Methoden [34, 163, 115], die den Komplex auf die Interaktionsdy-



(a) Die beste gemeinsame Substruktur der 5Å-Umgebungen von ADP (17 Residuen auf 1W2D.A und 18 Residuen auf 1ZAR.A) besteht aus 14 Residuen mit einer RMSD von 1.378Å, einer Identität von 50.0% und der gemessenen Ähnlichkeit von 87.72%.



(b) Die unmittelbaren Umgebungen von ADP sind in einem gemeinsamen, diskontinuierlichen Faltungsmotiv verankert. Das Motiv besteht aus etwa 60 Residuen mit einer RMSD von 1.648Å, einer Identität von 21.67% und der gemessenen Ähnlichkeit von 76.98%.



(c) Die Strukturen 1W2D.A (Human Inositol Trisphosphate 3-Kinase) und 1ZAR.A (Archaeoglobus Fulgidus Rio2 Kinase) gelten als nicht homolog. Beide Strukturen sind im Komplex mit ADP aufgeklärt und verfügen über eine homologe Funktion (Kinase). Die beste gemeinsame Substruktur besteht aus 105 Aminosäuren mit einer RMSD von 2.453Å, einer Identität von 19.05% und der gemessenen Ähnlichkeit von 63.28%. Gemessen an den Größen der beiden Strukturen (1W2D.A: 265 Residuen; 1ZAR.A: 267 Residuen) beläuft sich ihre Gesamtähnlichkeit auf 25.07%. Die beiden ähnlichen, jedoch nicht identischen Wechselwirkungsstellen haben die analoge Funktion.

Abbildung 2: Analoge Funktion.

namik und Bindungsaffinität untersuchen, da die Ausrichtung des Liganden am Mimotop bereits bekannt ist.

Mit der wachsenden Größe einer Wechselwirkungsstelle steigt auch ihr Potential für die Multispezifität [72]. Umgekehrt, können unterschiedliche Wechselwirkungsstellen mit dem gleichen Ligand interagieren. Analog zu der Unschärfe in der Abhängigkeit der Faltungsmuster von den Sequenzen, existiert eine Unschärfe in der Abhängigkeit der räumlichen Rückgratkonformation einer Wechselwirkungsstelle von der räumlichen Konformation der mit dem Liganden interagierenden aktiven Gruppen der Aminosäurenseitenketten. Die Konformation der Hotspot-Seitenketten ist somit nicht an eine einzigartige Konformation des Rückgrats gebunden. Diese Unschärfe lässt die Schlüssel-Schloss-Hypothese von Emil Fischer (1894) etwas verschwommen erscheinen, sodass figurativ ausgedrückt, die gleichen Schlösser sich von unterschiedlichen Schlüsseln und die unterschiedlichen Schlösser sich von gleichen Schlüsseln öffnen lassen. Positiver Effekt dabei ist ein gewisser Spielraum in der Modellierung des Schlosses. Abb. 2 zeigt zwei als nicht homolog geltenden Kinasen. Die Epitope, als 5$\AA$-Umgebungen von ADP mit 17 und 18 Aminosäuren, haben gemeinsame Substrukturen aus 14 Aminosäuren mit einer Identität von 50.0%, die insgesamt eine sehr ähnliche Oberflächentopologie besitzen (Abb. 2a). Gemessene Ähnlichkeit der beiden Substrukturen liegt bei 87.72%. Getragen werden die beiden Epitope von einem relativ ähnlichen Faltungsmuster (Abb. 2b), mit der Identität von 21.67% bei 60 Aminosäuren und der gemessenen Ähnlichkeit von 76.98%. Gemessen an den Größen der beiden Strukturen beläuft sich ihre Gesamtähnlichkeit auf 25.07%, die die gemeinsame Kinaseaktivität bzw. figurativ ausgedrückt das ADP-Schloss ausmacht (Abb. 2c).

Der Verlust von wenigen Hotspot-Residuen innerhalb der Wechselwirkungsstelle kann zu drastischen Veränderungen der Bindungsaffinität während der Interaktion führen [43]. Zusätzlich kann die Mutation der Residuen außerhalb der Wechselwirkungsstelle Veränderungen des elektrostatischen Potentials nach sich ziehen, das sowohl die gegenseitige Ausrichtung der Interaktionspartner vor der Interaktion beeinflusst als auch eine stabilisierende Wirkung auf den Protein-Ligand-Komplex während der Interaktion ausübt [157]. Ein Mimotop muss also unter der Berücksichtigung dieser beiden Gesichtspunkte designt werden. Allerdings garantiert das Design eines Mimotops nicht, dass neben der transplantierten Funktion noch weitere Funktionen erzeugt werden, die ihren Einsatz am Bestimmungsort sowohl zusätzlich begünstigen als auch gänzlich ausschließen können. Während die Antigenität eines Mimotops sich lediglich auf seine Reaktionsfähigkeit mit einem Antikörper bezieht, hängt seine Immunogenität von vielen komplexen Interaktionen mit unterschiedlichen Elementen des Immunsystems seines Wirtes ab [139]. Solange die entsprechenden zellulären und regulatorischen Mechanismen nicht vollständig verstanden sind, bleibt das Design der Mimotope ein äußerst komplexes Unterfangen.

In diesem Prozess beschäftigt sich EpitopeMatch mit der Auswertung der Ähnlichkeit der Epitope, mit der Suche nach den passenden Leitstrukturen und mit der Transplantation der Wechselwirkungsstellen auf die Leitstrukturen, und somit mit dem Design von Rohmimotopen.

1.2 VOM ALIGNMENT ZUM VERGLEICH

Der Fachausdruck *Alignment* in der Terminologie der Bioinformatik impliziert eine gerichtete Zuordnung der Aminosäuren zweier Proteine unter der Berücksichtigung ihrer Sequenzordnung bzw. ihrer Syntheserichtung. Während die Berücksichtigung der Kontinuität im Rahmen von Sequenzalignment eine Voraussetzung ist, decken die kontinuierlichen Align-

ments im Rahmen von Strukturalignment lediglich die kleinere Teilmenge aller möglichen Alignmentkombinationen zweier Strukturen ab, wobei die deutlich größere Teilmenge die Menge aller diskontinuierlichen Zuordnungen ist (Abs. 2.2.2, Abb. 3, Tab. 1). Somit dehnt sich der Begriff *Alignment* in Verbindung mit Struktur über die Ausrichtung an der Sequenzordnung hinaus und bedeutet in erster Linie den *Vergleich* auf der Basis von Geometrie. Erst die Verfügbarkeit der dreidimensionalen Daten, und somit die Kenntnis von der räumlichen Anordnung der Aminosäuren, ermöglicht die Erkennung, Charakterisierung und Klassifizierung von konservierten Faltungs-, Architektur- und Bindungsmotiven. Anhand der gemessenen Ähnlichkeit der Strukturen und Substrukturen lassen sich mit den Methoden des Data-Mining die funktionalen und die entfernten evolutionären Zusammenhänge ermitteln.

Neben Strukturmodellierung, Docking und Protein-Engineering ist das Structure Comparison and Alignment (SCA) die am häufigsten verwendete Basisoperation [182] und gilt als stärkstes Werkzeug für die Lösung des Puzzles der funktionalen Proteinbeziehungen (Gene Ontology (GO)) in den und zwischen den Proteomen. Performance, Qualität und Bandbreite eines Strukturalignmentalgorithmus stehen im unmittelbaren Zusammenhang mit der Wahl der Auflösung der vorliegenden Strukturturinformation. Unter der "Wahl der Auflösung" ist die Extraktion der Struktureigenschaften (Feature Extraction (FE)) [45] zu verstehen, die sowohl miteinander verglichen als auch nach ihrer Ähnlichkeit bzw. Unähnlichkeit kategorisiert werden können. Im Allgemeinen sind alle Substrukturen aller bekannten Strukturen miteinander vergleichbar, was die Strukturalignmentalgorithmen bezüglich der Erkennung von wiederkehrenden Substrukturen vor eine immense Herausforderung stellt. Quasi jedes Strukturpaar verfügt über Ähnlichkeiten, die auf dem durch die physikalischen und chemischen Gesetzmäßigkeiten eingeschränkten Faltungsraum der Proteine und auf ihren evolutionären Beziehungen beruhen [91]. Die zweite Herausforderung ist es, zwischen den Ähnlichkeiten zu unterscheiden. Jede Substruktur der Proteine enthält zwangsläufig die gesamte Information über ihre einzelnen Komponenten (Aminosäuren): die relative räumliche Konformation; die Typen mit den jeweiligen physiko-chemischen Eigenschaften; die Zugehörigkeit zu den Sekundärstrukturen und der Solvent Accessible Surface (SAS); die Torsionswinkel des Rückgrats und der Seitenketten; das elektrostatische Potential; das Volumen. Je vollständiger die Eigenschaften einer Substruktur beschrieben sind, desto genauer bewertet die entsprechende Objective Function (OF) ihre Ähnlichkeit zu einer anderen Substruktur. Die ausgewählten Eigenschaften werden in Form von unterschiedlichen Datenstrukturen (Listen, (Hash-)Tabellen, Matrizen, Bäumen, Graphen, etc.) repräsentiert, die den Einsatz von allgemeinen effizienten algorithmischen Methoden (Dynamic Programming (DP), Geometric Hashing (GH), Monte Carlo (MC), Genetic Algorithm (GA), Combinatorial Extension (CE), Clique Detection (CD), etc.) erlauben. Die Strategien zur Bewältigung des Strukturalignmentproblems verzweigen sich generell in zwei Richtungen: Präprozessierung der Strukturinformation im Hinblick auf die ausgewählten Eigenschaften und Datenstrukturen, um den Kombinationsraum drastisch zu verringern und somit das Vergleichen zu beschleunigen; oder der Vergleich on-the-fly, der in der Regel eine heuristische Lösung für das Umgehen des Non-deterministic Polynomial-time (NP)-harten, kombinatorischen Problems [94] darstellt. Die meisten Algorithmen bestehen aus zwei oder mehr Stufen, im deren Verlauf das anfangs ermittelte initiale Alignment zu einem globalen Alignment ausgebaut wird, wobei die einzelnen Stufen Implementierungen der jeweiligen Strategie sein können. Die meisten OFs konzentrieren sich lediglich auf die Ausbalancierung der zu minimierenden geometrischen Abweichung und der zu maximierenden Alignmentgröße, indem die RMSD (Gl. 31) nach der Superpositionierung der Substrukturen ins Verhältnis zu der Anzahl und der Identität der korrespondierenden Elemente gebracht wird. Diese Bewertungsweise, und die Orientierung an den kontinuierlichen Fragmenten, erlau-

ben die Klassifizierung der Strukturen nach ihren Faltungsmustern, Domänen, Topologien, Architekturen und Familien (Structural Classification Of Proteins (SCOP) [118, 8, 9], Class Architecture Topology Homologous superfamily (CATH) [128, 130, 162], Family of Structurally Similar Proteins (FSSP) [70, 66, 67, 68]) [126]. Die meisten Algorithmen aus dieser Klasse (Abs. 1.2.1) beherrschen die Erkennung der Faltungsmuster unabhängig von den Insertion/Deletion (ID)s [36]. Nur wenige beherrschen die Flexibilität (beträchtliche Konformationsänderungen als Folge der Domänenverschiebung). Keines kann mit dem Problem der Circular Permutation (CP)s [148] umgehen, die zu einer Veränderung der Topologie, d.h der Secondary Structure Element (SSE)-Konnektivität bezüglich der Syntheserichtung führen.

Das zugrunde liegende Problem des Strukturalignments ist das klassische Problem der computergestützten Geometrie - das Vergleichen von Punktwolken im Raum [91]. Abgesehen von den kovalenten Bindungen des Rückgrats und der Seitenketten sind die Punktwolkenmuster unregelmäßig. Erschwerend kommen die Musterverzerrungen (Induced-Fit) hinzu, die durch die Flexibilität der Seitenketten an der Oberfläche und ggf. größere Konformationsänderungen durch Domänenverschiebungen hervorgerufen werden. Angefangen mit dem CP-Problem und der Verschiebung der Betrachtung von den Faltungsmustern auf die diskontinuierlichen geometrischen Muster der Proteinoberflächen, muss man sich von der Gruppierung der Elemente unter der Berücksichtigung der Sequenzordnung trennen. Mit anderen Worten, die alleinige Betrachtung der Cα-Beziehungen ist für die Erkennung der Bindungsstellen nicht mehr ausreichend. Die drei gängigsten Lösungsansätze für dieses Problem sind die Betrachtung der räumlichen Residuenmuster, der räumlichen Muster der aktiven Gruppen der Residuen (Hotspots) und der gerasterten oder der triangulierten Oberflächen. Die Techniken basieren entweder auf den Distanzen oder auf den Koordinaten, wobei die letzteren die Koordinaten der Gitterzellen mit den Punkten der Muster sind. Neben den geometrischen Musterverzerrungen existiert ein weiteres Problem - die Multimodalität der Bindung, bei der der gleiche Ligand sowohl in den unterschiedlichen Modi an die gleiche Bindungsstelle als auch im gleichen Modus an die unterschiedlichen Bindungsstellen binden kann. Die Oberflächentriangulierung erlaubt die möglichst genaue Beschreibung der Bindungsstellenoberflächen, die Einbeziehung des Electrostatic Potential (EP) und trennt sich von dem tragenden Faltungsmuster. Der Vergleich der triangulierten Oberflächen reagiert jedoch empfindlich auf die Konformationsänderungen und ist aufgrund der viel höheren Anzahl von Vertices (gegenüber der Anzahl der Oberflächenatome) deutlich rechenintensiver. Die auf dem Vergleich von Residuen- oder Hotspot-Mustern basierenden Methoden erlauben eine größere Konformationsfreiheit und die Definition von Konsensus-Mustern [98], die dem Problem der Multimodalität der Bindung entgegenwirken. Darüber hinaus erlauben die auf den Residenmustern basierenden Methoden, die mehr an die Faltungsmuster gebunden sind, die Transplantation der Bindungsstellen durch die Mutation der entsprechenden Residuen.

Die nächsten Abschnitte fassen die bekanntesten Algorithmen aus der jeweiligen SCA-Rubrik in Form einer hierarchischen Übersicht zusammen.

1.2.1 *Kontinuierliches Alignment*

Die kontinuierlichen Methoden stützen sich auf die topologischen Eigenschaften der Proteine: die konsekutiven Rückgratfragmente oder die SSEs ([142]).

DALI (Distance matrix ALIgnment) [65] nutzt die rotations- und translationsinvariante Repräsentation der Strukturen in Form von Distanzmatrizen (Abs. 2.2.3, Abb. 4), die

in Hexapeptid-Submatrizen fragmentiert werden. Jede Submatrix stellt ein Cα-Cα-Kontaktmuster dar. Die Fragmente zweier Strukturen werden verglichen und zu den größeren überlappenden Segmenten assembliert. Die Assemblierung der Fragmente zu den Segmenten wird durch eine MC-Methode optimiert, die den Suchraum, mit den gelegentlichen Ausreißern in die nicht optimalen Bereiche, nach Zufallsprinzip durchwandert. Die Methode generiert mehrere Lösungen. Gaps sind zulässig. Wie alle heuristischen Methoden (Abs. 2.2) garantieren die MC-Methoden keine optimale Lösung. Die auf der Basis von DALI aufgebaute Datenbank FSSP [70, 66, 67, 68] erlaubt die automatische Klassifizierung der Proteinstrukturen nach ihren Faltungsmotiven durch hierarchisches Clustern der paarweisen Alignments. Die Datenbank Homology-derived StructureS of Proteins (HSSP) [150, 69] assoziiert die Sequenzen mit den Strukturen und impliziert 3D-Modelle der bekannten Sequenzen über ihre Homologie zu den Sequenzen der bekannten Strukturen (Abs. 1.1.2.1).

CE (Combinatorial Extension) [158, 159] fragmentiert die Strukturen in Oligo- bzw. Polypeptide ($4 \leq m \leq 36$), wobei die Oktapeptide sich als ausbalanciert bezüglich der Sensitivität und der Performance erwiesen haben. Fragmente werden anhand der lokalen Geometrie aligniert und die so entstehenden Aligned Fragment Pair (AFP)s zu einem optimalen, konsekutiven Pfad verbunden, wobei Gaps zulässig sind. Ähnlichkeit wird anhand der RMSD gemessen. Als Signifikanzmaß wird ein Z-Score verwendet, der auf der Wahrscheinlichkeit basiert, ein Alignment der gleichen Länge beim Vergleich zweier zufällig ausgewählten Strukturen zu finden. CE ist seit der Erweiterung auf die Erkennung der zirkulären Permutationen [135] ein integraler Bestandteil der PDB.

MATALIGN (MATrix ALIGNment) [15] repräsentiert die Strukturen in Form von Distanzmatrizen und betrachten jede Reihe der symmetrischen Matrix als ein Distanzen-Profil. Die Reihen werden mit der DP-Methode [121] aligniert. Auf die resultierende Scoringmatrix wird erneut die DP-Methode angewendet um die korrespondierenden Residuenpaare der initialen Alignments zu ermitteln. Mittels einer Objective Scoring Function (OSF) aus [4], die eine Balance zwischen RMSD und Alignmentgröße darstellt, wird das initiale Alignment iterativ optimiert.

SHEBA (Structural Homology by Environment-Based Alignment) [80, 81] erstellt Umgebungsprofile für jede Aminosäure einer Struktur, die sich aus Komponenten Sequenzhomolgie, SSE-Typ, Solvent Accessibility (SA) und Polarität der Umgebung [27] zusammensetzt. Die Profile zweier Strukturen werden anhand der DP-Methode aligniert und das daraus resultierende initiale Alignment superpositioniert [83, 84]. Das Alignment wird optimiert bis die maximale Anzahl der Cα-Distanzen $< 3.5\mathring{A}$ zwischen den korrespondierenden Residuen erreicht ist.

MATRAS (MArkovian TRAnsition of protein Structure) [86, 87] nutzt eine hierarchische Alignmentstrategie: zuerst clustert es die SSEs anhand der SSE-Scores nach der Methode von [114]; das "grobe" SSE-Alignment wird dann anhand einer DP-Methode unter Verwendung eines lokalen Umgebungs- und eines Distanzen-Scores verfeinert. Die Ähnlichkeit zweier alignierten Strukturen wird nach dem Prinzip des Markovian Transition Model of Evolution (MTME) (ähnlich zu der Point Accepted Mutation (PAM)-Matrix [42]) bewertet, wobei die Wahrscheinlichkeit der Transition der Target-Struktur (TS) nach Query-Struktur (QS) im Prozess der Evolution im Verhältnis zu der Wahrscheinlichkeit eines zufälligen Erscheinens der QS gemessen wird.

VAST (Vector Alignment Search Tool) [108] segmentiert die Strukturen in SSEs, von denen jede als ein Knoten im Rahmen eines graphentheoretischen Ansatzes repräsentiert wird. Die Kanten des Graphen basieren auf dem Typ, der relativen Orientierung (Winkel) und der Konnektivität (Distanzen) der SSEs. Für die signifikante Beschreibung der Strukturähnlichkeit werden Homologous Core structure overlap Score (HCS), Loop Hausdorff Measure (LHM) und Gapped Structural Alignment Score (GSAS) [109] verwendet. Auf VAST basierende Molecular Modeling DataBase (MMDB) [110] erlaubt neben der Suche nach den ähnlichen Strukturen die Suche nach ähnlichen makromolekularen Komplexen.

GANGSTA (Genetic Algorithm for Nonsequential, Gapped protein STructure Alignment) [93, 56] definiert eine Kontaktmatrix auf der Ebene von Residuen und auf der Ebene von SSEs. Zwei Residuen stehen im Kontakt, wenn die Distanz zwischen ihren Cα-Atomen $< 11\,\mathring{A}$ ist [21]. Zwei SSEs stehen im Kontakt, wenn sie über mindestens ein Kontakt auf der Residuenebene verfügen. Die SSEs werden je nach Typ [82] und Länge als Knoten eines Graphen repräsentiert, mit der Anzahl der Residuenkontakte und der relativen Orientierung zwischen den SSE-Paaren [90] als Kanten. Bezüglich der Topologie können die SSE-Paare parallel, antiparallel oder gemischt verlaufen. Mittels einer GA-Methode wird nach den größtmöglichen Subgraphen bzw. Individuen gesucht. Die korrespondierenden SSEs müssen vom selben Typ sein und um nicht mehr als 10 Residuen in ihrer Länge voneinander abweichen. Ein Subgraph mit den korrespondierenden SSEs wird als ein Gen betrachtet. Ein Allel des jeweiligen Gens beschreibt seine unterschiedlichen Mutationen. Neue Generationen werden durch die Mutation und den Austausch der Gene erzeugt. Eine OF entscheidet über die Fitness der neuen Individuen. Daraus resultierendes Strukturalignment auf der SSE-Ebene wird durch das Verschieben der korrespondierenden SSEs im Hinblick auf die Maximierung der Kontakte auf der Residuenebene und die Minimierung des auf der RMSD basierenden Scores optimiert.

WHAT-IF [176] fragmentiert die Strukturen in Polypeptide ($10 \leq m \leq 15$), wobei die kürzeren Fragmente zu dem Problem der kombinatorischen Explosion (Abs. 2.2.2) und die längeren Fragmente zu einem Mangel an richtig-positiven Zuordnungen der Fragmente führen. Die überlagerten Fragmentpaare werden nach der RMSD geclustert. Die Cluster repräsentieren die gemeinsamen Substrukturen.

1.2.1.1 *Flexibles Alignment*

FATCAT (Flexible structure AlignmenT by Chaining AFPs with Twists) [184] überlagert konsekutiven Fragmente (z.B. Oktapeptide) zweier Strukturen. 2 Fragmente gelten als kompatibel, wenn die RMSD ihrer Distanzmatrizen signifikant klein ist. 2 AFPs gelten als kompatibel, wenn ihre Überlagerungen zu ähnlichen Überlagerungen der beiden gesamten Strukturen führen. Das gesamte Alignment ist eine Kette von konsekutiven AFPs, die durch Hinge-Twists voneinander getrennt sind. Optimale Verkettung der AFPs wird anhand einer DP-Methode ermittelt.

FLEXPROT (alignment of FLEXible PROTein structures) [153, 154, 155] erweitert jedes Atompaar der QS und TS iterativ um weitere Atompaare nach links und rechts entlang des Rückgrats bis die korrespondierenden, konsekutiven Fragmente die Größe von 12 Aminosäuren erreicht haben oder die RMSD $> 3\mathring{A}$ geworden ist. Repräsentiert die kongruenten Fragmentpaare in Form von einem gerichteten und gewichteten azyklischen

Graphen mit den Fragmentpaaren als Knoten und hinge-Regionen als Kanten, und sucht nach den kürzesten Pfaden unter Berücksichtigung der Sequenzordnung. Die Pfade werden nach Größe und RMSD sortiert und ausgegeben. Die Hinge-Bending-Winkel und Hinge-Twist-Torsionswinkel zwischen den starren Körpern (rigid body) und den Hinge-Regionen werden nach [111] als Winkel zwischen den Zentroiden des ersten starren Körpers, der Hinge-Region und des zweiten starren Körpers bzw. als Torsionswinkel zwischen dem Zentroid des ersten starren Körpers, dem an die Hinge-Region angrenzenden Cα-Atom des ersten starren Körpers, dem an die Hinge-Region angrenzenden Cα-Atom des zweiten starren Körpers und dem Zentroid des zweiten starren Körpers berechnet.

1.2.2 *Diskontinuierlicher Vergleich*

CA-MATCH [127, 17] nutzt eine GH-Methode: beginnt mit einer Präprozessierung der geometrischen Daten in eine rotations- und translationsinvariante Repräsentierung in Form von orthonormalen Referenzframes. Ein Referenz-Cα-Atompaar bildet jeweils ein Dreieck zu jedem anderen Cα-Atom der Struktur. Das Distanzentriplet eines Dreiecks bildet einen eindeutigen Hashwert, der in einer Hashtabelle gespeichert wird. Die Anzahl der Dreiecke wird durch Distanzen-Thresholds begrenzt. Die Atome der korrespondierenden Dreiecke des Modells und des Targets werden gezählt und anschließend geclustert. Die so erhaltenen Seed Match (SM)s werden mittels einer Heuristik zu einem diskontinuierlichen Alignment zusammengesetzt und anhand der Größe des Alignments und der RMSD bewertet.

SCALI (Structural Core ALIgnment) [187] findet topologieunabhängigen Architekturmuster als Ähnlichkeitskerne in drei Schritten: generiert lokale Sequenz-/Struktur-Alignments der kontinuierlichen Fragmente (> 5), indem die Zuordnung jedes Fragments der QS jedem Fragment der TS als Summe der Wahrscheinlichkeiten für die Markov-Zustände nach Hidden Markov Model for local STRucture (HMMSTR) [32] bewertet wird, wobei jeder Markov-Zustand einem I-sites-Motiv [31] entspricht, der Informationen über die bevorzugten Aminosäuren und Torsionswinkel enthält; superpositioniert die Fragmente anhand ihrer Distanzmatrizen, wählt die besten Matches als initiale Alignments und erweitert diese mit einer Breadth-First Tree Search (BFTS)-Methode bis keine Fragmente mehr hinzugefügt werden können; optimiert die besten globalen Alignments nach RMSD durch Entfernen/Hinzufügen von alignierten Blöcken wenn alle/keine Distanzdifferenzen $> 9\mathring{A}$ und Torsionswinkeldifferenzen $> 100°$ sind.

SARF (Spatial ARrangement of backbone Fragments) [5, 4, 3] fragmentiert die Strukturen in Pentapeptide, ordnet diesen die SSE-Typen (α, β) zu, benutzt die Winkel und die Distanzen zwischen SSE-Achsen zur Filterung nach SSE-Paaren, die mit einer kleinen RMSD überlagert werden können. Zum Umgehen der kombinatorischen Explosion (Abs. 2.2.2) beim Aufbau der größtmöglichen SSE-Ensembles werden nur die benachbarten ($< 25\mathring{A}$) SSEs berücksichtigt. Anschließend wird das größte Ensemble um weitere SSEs erweitert um die Alignmentgröße zu maximieren, wobei die RMSD gleichzeitig zu minimieren ist.

1.2.2.1 *Flexibler Vergleich*

FLEXSNAP (FLEXible Non-Sequential Protein structure Alignment) [145] extrahiert konsekutiven AFPs mit der Länge ≥ 3, indem die Fragmente der QS entlang des Rückgrats

der TS überlagert werden und nach der RMSD bewertet werden, die vereinfacht berechnet wird (ohne Anwendung der Transformation auf die Koordinaten). Anschließend werden Sets von nicht überlappenden AFPs gebildet, sodass die Größe der Sets maximiert, die RMSD der Sets minimiert und die Anzahl der Hinge-Regionen und Gaps ebenfalls minimiert wird. Statt eines graphentheoretischen Ansatzes (siehe FlexProt) wird für die Bildung der AFP-Sets eine greedy-Methode verwendet, bei der die längsten AFPs als Startpunkte für eine iterative Erweiterung mit neuen AFPs dienen, wobei die Sets über ihre Größe, ihre RMSD, die Anzahl der Hinge-Regionen und der Gaps ausbalanciert wird.

DEDAL (DEscriptor Defined Alignment) [41] definiert lokale Deskriptoren als lokale Umgebungen der Residuen. Ein Deskriptor enthält ein zentrales Residuum, die benachbarten Residuen (wenn die Distanzen zwischen ihren Cα-Atomen und den geometrischen Zentren ihrer Seitenketten (Im Fall von GLY Cα-Atom) zu den selbigen des zentralen Residuums $\leq 6.5\mathring{A}$ bzw. $\leq 8\mathring{A}$ sind) und jeweils zwei Residuen in beide Richtungen des Rückgrats um jedes benachbarte Residuum. Auf diese Weise entstehen Kombinationen von benachbarten konsekutiven Fragmenten. Diese werden miteinander überlagert, wobei die relative Verteilung der benachbarten Residuen um die entsprechenden zentralen Residuen möglichst gleich ($RMSD \leq 2.5\mathring{A}$) und mindestens die Hälfte aller benachbarten Residuen vorhanden sein muss. Jedes alignierte Deskriptorpaar ist ein partielles Strukturalignment. Diese können miteinander kombiniert werden, wenn sie nach der gemeinsamen Überlagerung konsistent sind (kleine RMSD). Das Konzept wird auf einen ungerichteten Graphen projiziert, mit Deskriptorpaaren als Knoten und ihrer Konsistenz als Kanten. Mit einer Branch and Bound (BB)-Methode wird ein Entscheidungsbaum solange traversiert, bis die größten Cliquen [28] gefunden ist.

1.2.2.2 *Multipler Vergleich*

MULTIPROT (simultaneous alignment of MULTIple PROTein Structures) [156] bietet eine heuristische Lösung des Multiple STructural Alignment (MSTA)-Problems an. Nutzt FlexProt-Algorithmus und aligniert alle kontinuierlichen Fragmente (≥ 3) der $m-1$ Strukturen mit den Fragmenten einer zentralen (pivot) Struktur. Diese Vorgehensweise wird für jede Struktur wiederholt, sodass jede Struktur ein Mal als zentrale Struktur fungiert. Nach jeder Iteration pro zentrale Struktur werden die zu einem bestimmten konsekutiven Abschnitt der zentralen Struktur gehörenden Fragmente der restlichen Strukturen gruppiert (cuts). Die gruppierten multiplen AFPs werden zu den globalen multiplen Alignments kombiniert, indem nach den größten kongruenten multiplen AFPs, die in möglichst vielen Strukturen enthalten sind, gesucht wird. Die entstehenden Lösungen, bzw. die geometrischen Kerne, werden nach der multiplen RMSD bewertet, indem ein RMSD-Mittelwert aller Strukturen zu einer ausgewählten zentralen Struktur berechnet wird.

1.2.2.3 *Bindungsseiten*

ASSAM (Amino acid pattern Search for Substructures And Motifs) [12, 167, 13, 119] spezialisiert sich auf die Suche nach den räumlichen Mustern der Aminosäurenseitenketten. Die in Form von Vektoren repräsentierten Residuen bilden die Knoten und die Distanzen zwischen den Vektorkoordinaten die Kanten eines Graphen, der mit dem Subgraph-Isomorphismus-Algorithmus [172] nach Mustern durchsucht wird. Jeder Vektor charakterisiert mit seinem Start- (S), End- (E) und Mittelpunkt (M) die

physiko-chemischen Eigenschaften einer Aminosäure, wobei einer von diesen Punkten (Pseudoatome) ein Schlüsselpunkt (K) ist, der die wichtigste Eigenschaft eines Residuums hervorhebt. Eine Kante wird somit über fünf Distanzen MM, KK, SS, SE, ES und EE beschrieben und impliziert somit auch die relative räumliche Ausrichtung der Residuen. Zusätzlich werden die Knoten mit SSE-, SA- und Disulfide Bridge (DB)-Information ausgestattet.

CLICK [123, 124] extrahiert Residueneigenschaften in Form von Koordinaten der repräsentativen Atome, SSE, SA und Residue Depth (RD) [35]. Definiert Cliquen aus 3-7 repräsentativen Atomen als lokale Fragmente und vergleicht diese Mittels Überlagerung. Auf der Basis der Äquivalenzen der kongruenten Cliquen werden globale Alignments berechnet, indem ihre Größe maximiert und die RMSD minimiert wird, wobei auch Alternativalignments ausgegeben werden.

TOPOFIT [74, 1, 103] tesselliert die Strukturen in Tetraederzellen, deren Ecken Cα-Atome sind und die Kanten so ausgewählt sind, dass die 4 Cα-Atome immer die nächsten Nachbarn sind. Die benachbarten Ecken werden in Form eines Graphen gespeichert und definieren somit die Nachbarschaft der Tetraeder. Die benachbarten Tetraeder bilden somit ein lückenloses Volumen. Die Tetraeder werden nach Form, Volumen und Rückgrattopologie klassifiziert. Die kongruenten Tetraeder (die Saat) der QS und der TS werden anhand der Überlagerung der ähnlichen Tetraeder ermittelt. Die Saat wird systematisch um die neuen kongruenten Tetraeder erweitert, sodass nach den größten gemeinsamen Subgraphen mit dem größtmöglichen Volumen bei einer kleinstmöglichen RMSD gesucht wird.

SITEENGINES [160, 161] fasst Aminosäurenatome mit ähnlichen physiko-chemischen Eigenschaften zu Pseudozentren zusammen, von denen jedes in Form eines Punktes im Raum repräsentiert wird. Jedem Pseudozentrum wird eine für die Protein-Ligand-Interaktion wichtige Eigenschaft nach [149] zugewiesen: hydrogen-bond DOnor (DO), hydrogen-bond ACceptor (AC), mixed Donor/Acceptor (DA), hydrophobic ALiphatic (AL) und aromatic PI contact (PI). Es werden nur die Pseudozentren berücksichtigt, die sich an der SAS [39, 40] befinden und nicht weiter als 4$\AA$ vom Liganden entfernt sind. Nutzt die GH-Methode (siehe Cα-match), wobei die Hashwerte der Dreiecke aus den Distanzen zwischen den Pseudozentren und ihren Eigenschaften gebildet werden. Transformationsdaten aus der Überlagerung der kongruenten Dreiecke werden für die Überlagerung der QS und TS verwendet. Nach der Überlagerung der QS und TS werden die Ähnlichkeiten der lokalen Umgebungen um die Pseudozentren gemessen.

TESS (TEmplate Search and Superposition) [177] wendet die GH-Methode an und stellt jede Aminosäure einer Struktur in Form eines orthonormalen Seitenketten-Referenzframes dar, mit jeweils drei repräsentativen Atomen. Ein Seitenkettenatom wird auf den Ursprung des Koordinatensystems transformiert, sodass die anderen beiden in positive x-Richtung zeigen. Die Transformation wird auf alle Aminosäuren aus der 18.0$\AA$-Umgebung der Referenzaminosäure angewendet. Der Referenzframe wird in ein 1.0$\AA$-Gitter gelegt und die Positionen der Gitterzellen mit mindestens einem Atom aus der Umgebung gespeichert, wobei auch der Typ der jeweiligen Aminosäure notiert wird. Auf diese Weise wird die gesamte PDB präprozessiert, sodass bei der Suche nur noch das Query präprozessiert werden muss. Bei der Suche werden die Inhalte der Gitterzellen miteinander verglichen, wobei die Genauigkeit von der Anzahl der zu durchsuchenden Nachbarzellen abhängt.

JESS (Jonathan's template Search and Superposition) [20] erzeugt Templates mit einer gewissen Anzahl von Atomen, wobei jedem Atom geometrische Eigenschaften zugewiesen werden. Anstatt eines Gitters (TESS) werden die räumlichen Muster mittels eines k-d-Baums [22] kodiert. Der k-d-Baum wird mit einem Backtracking-Algorithmus nach Lösungen bzw. Teillösungen durchsucht.

PROFUNC [101] definiert 3D-Templates als spezifische Konformationen von wenigen Residuen (2 bis 5) der aktiven Zentren der Enzyme [134], Liganden- und DNA-Bindungsresiduen. Die zwei letzteren werden von den Komplexen (Holo-Strukturen) aus der PDB abgeleitet. Bei Apo-Strukturen werden die sogenannten reverse-Templates generiert, indem die Struktur in viele 3-Residuen-Templates, wobei die Residuen benachbart ($\leq 5\mathring{A}$) sein müssen, gespalten wird. Die Templates werden anhand der Anzahl der Ionen-, Wasserstoffbrücken-Bindungen und hydrophoben Kontakte charakterisiert und nach der Methode von JESS gespeichert und durchsucht.

LIGSITE [60] platziert eine Struktur in ein 3D-Gitter und scannt die Zellen entlang der Achsen und der Diagonalen nach Protein-Lösung-Protein-Ereignissen (PSP). Alle Zellen mit SAS-Atomen werden zwischen PSP-Werten 0 (kein PSP-Ereignis) und 7 (tief vergraben, zwischen anderen Zellen mit SAS-Atomen) markiert. LIGSITEcsc [60] erweitert diesen Ansatz, indem die SAS nach [39, 40] berechnet wird und die PSP-Ereignisse durch Oberfläche-Lösung-Oberfläche-Ereignisse (SSS) ersetzt werden, sodass nur noch die Zellen mit den Oberflächenvertices in Betracht gezogen werden. Auf der Grundlage dieser Repräsentierung basiert die Datenbank CavBase [149]. Die gespeicherten Bindungsseiten werden um die Pseudozentren ergänzt, von denen jedes physiko-chemische Eigenschaften (Donor, Akzeptor, aliphatisch, aromatisch, etc.) zugewiesen bekommt. Das Matchen der Bindungsseiten erfolgt mittels Cliquen-Erkennung [28].

EF-SITE (Electrostatic surface of Functional-SITE in proteins) [88, 89] trianguliert die Strukturoberfläche [39, 40], ergänzt die Vertices mit den Werten des EP [120] und der minimalen und maximalen Krümmung um den jeweiligen Vertex und definiert diese als Konten eines Graphen. Zwei Vertices zweier Strukturen sind dann korrespondierend, wenn die Unterschiede im EP und in den Krümmungen minimal sind. Auf diese Weise wird die Anzahl der korrespondierenden Vertices drastisch reduziert. Zwei Vertex-Paare gelten als über eine Kante miteinander Verbunden, wenn die Vertex-Distanzen der jeweiligen Struktur $< 1.5\mathring{A}$ sind. Das Matchen der triangulierten Oberflächen erfolgt mittels Cliquen-Erkennung [28]. In der neusten Version [117] wird die Cliquen-Erkennung aus Gründen der Performance durch GH ersetzt.

EPITOPEMATCH [77] Abs. 2.1 ist eine detaillierte Zusammenfassung der neuen EPITOPEMATCH-Version.

In diesem Rahmen beschäftigt sich EPITOPEMATCH mit dem diskontinuierlichen Strukturalignment an sich, wobei das kontinuierliche Strukturalignment lediglich als Nebenprodukt betrachtet werden kann und keinen Einsatzschwerpunkt von EPITOPEMATCH darstellt.

Teil II

EPITOPEMATCH

Teil ii stellt ausführlich die Software EpitopeMatch vor, die im gegenwärtigen Entwicklungsstand ein Verfahren zur Identifikation von Leitstrukturen für die Transplantation gegebener Epitope ist. Der Abschnitt konzentriert sich insbesondere auf die Bewertung der Ähnlichkeit und die Anwendungsbandbreite.

2 ALGORITHMUS

2.1 ÜBERSICHT

Die primäre Aufgabe von EpitopeMatch (Abb. 79) ist die Erkennung, Bewertung und ggf. die Übertragung der Strukturähnlichkeit auf der Basis der nativen oder modellierten dreidimensionalen Konformationen der Aminosäuren. EpitopeMatch ist eine universelle Basisoperation aus dem Bereich SCA. Das Adjektiv "universell" betont die Einsatzbandbreite der Methode, die durch die Charakteristiken der Datenrepräsentation, der Heuristik und des Scoring, möglichst viele SCA-Facetten abzudecken versucht.

Der zugrunde liegende Vergleich der Punktwolken ist distanzenbasiert, sodass keine Präprozessierung der Koordinaten (Orthonormalisierung) durchgeführt wird. Die Punkte eines Musters können gruppiert werden. Die kleinste Gruppe ist ein Punkt, der entweder ein Atom, ein geometrisches Zentrum von N Atomen oder ein Oberflächenvertex sein kann. Eine Gruppe kann eine beliebige Kombination der Atome einer chemischen Komponente aus Chemical Component Dictionary (CCD) (z.B. Aminosäure), eine beliebige Kombination der geometrischen Zentren der Atomkombinationen einer chemischen Komponente oder z.B. die Vertex-Normale-Kombination der Oberfläche sein. Die Vermessung der Distanzen zwischen den Gruppen mit mehr als einem Punkt impliziert ihre relative räumliche Ausrichtung. Die korrespondierenden Elemente der Gruppen sind sinngemäß bzw. gemäß der Nomenklatur über mehrere Distanzmatrizen verteilt, von denen jede eine korrespondierende Elementebene darstellt (Abs. 2.2.3). Die Definition der Distanzmatrizen erfolgt anhand der Templates (Abb. 5). Jede Gruppe verfügt über physiko-chemische Eigenschaften (Abs. 2.2.4). Diese können in einem Ausdruck zusammengefasst und normalisiert werden. Auf dieser Grundlage wird für jedes mögliche Gruppenpaar die Substitutionsähnlichkeit ermittelt, die ebenfalls in Form von Templates (Abb. 7, Abb. 37) gespeichert wird. Die templateorientierte Definition der geometrischen und physiko-chemischen Eigenschaften gestattet eine flexible Kombination der einzelnen Facetten der Muster, auf die man sich in einem oder anderem Fall konzentrieren möchte.

Aus zwei Gruppen der QS mit den ähnlichen Distanzen zu zwei Gruppen der TS resultieren vier korrespondierenden Gruppenpaare, zwei in die parallele und zwei in die antiparallele (Synthese-)Richtung, wobei ein korrespondierendes Gruppenpaar die kleinste Common Substructure (CS) darstellt (Abs. 2.2.5). Im Prinzip kann jede Gruppe aus QS jeder Gruppe aus TS entsprechen, es sei denn, die Substitutionsähnlichkeit des entsprechenden Gruppenpaares ist gleich Null. Das Ziel ist die Kombination der korrespondierenden Gruppenpaare zu der BCS (Ähnlichkeitskern) und der Maximum Common Substructure (MCS) im Rahmen einer OF, die sowohl die geometrische als auch die physiko-chemische Ähnlichkeit in sich vereint und für jede CS gilt. Die richtig-positiven Gruppenpaare sind insbesondere im Fall von unterschiedlich großen Punktwolken stark verrauscht (Distanzenrauschen, Abb. 8a). Aus diesem Grund ist der einmalige Vergleich der Distanzmatrizen für das Herausfiltern der korrespondierenden richtig-positiven Gruppenpaare nicht möglich. EpitopeMatch kehrt das Problem um und entfernt die einfacher zu ermittelnden richtig-negativen Gruppenpaare so lange aus dem Gesamtbild, bis die Anzahl der verbliebenen Gruppenpaare der Anzahl der Gruppen der größeren Struktur entspricht und die Menge der zugeordneten Gruppen der größeren Struktur voneinander verschiedene Gruppen enthält. Dieser deter-

ministische Prozess der iterativen Rauschunterdrückung bildet die erste Algorithmusstufe und resultiert in einem SM - einer Menge der richtig- und ggf. falsch-positiven korrespondierenden Gruppenpaare (Abb. 8b), die in der Regel sowohl die mehrfach vorhandenen Muster als auch Teile der verzerrten Muster enthält. Der Ansatz der iterativen Rauschunterdrückung wurde bislang von keinem der bekannten Algorithmen verfolgt.

Die zweite Algorithmusstufe (Abs. 2.2.6) beschäftigt sich mit der Kombination der korrespondierenden Gruppenpaare aus SM zu einem initialen Alignment, das anschließend zu der besten gemeinsamen Substruktur (BCS) und der größten gemeinsamen Substruktur (MCS) optimiert wird, wobei die BCS gleichzeitig die MCS sein kann. Die korrespondierenden Gruppenpaare aus SM können als alignierten Fragmentpaare (AFPs) der Länge 1 bzw. als die kleinstmöglichen gemeinsamen Substrukturen (CSs) der Ordnung $k = 1$ verstanden werden. Jede gemeinsame Substruktur (CS) ist eine Kombination k-ter Ordnung. Die Ordnung k ist die Anzahl der korrespondierenden Gruppenpaare einer gemeinsamen Substruktur. Kombinatorisches Resampling der CSs erster Ordnung zu den CSs höherer Ordnungen anhand der CSs erster Ordnung ist von der Sequenzordnung völlig unabhängig und somit absolut diskontinuierlich. Wenn jede CS aus SM richtig-positiv ist (Idealfall), dann sind alle Kombinationen der Ordnungen $1 \leq k \leq |SM|$, die aus den CSs der Ordnung $k = 1$ gebildet werden können, ebenfalls richtig-positiv. Wenn jede dieser Kombinationen eine richtig-positive CS und somit eine potentielle Lösung des Matching-Problems ist, dann stellt sich die Frage, welche dieser Kombinationen bzw. Alternativalignments von Interesse ist? Die meisten Algorithmen begnügen sich mit der Antwort, dass lediglich eine CS mit der besten Balance zwischen der RMSD und der Größe interessant sei. In der Regel handelt es sich hierbei um den Ähnlichkeitskern (BCS). Die Antwort von EpitopeMatch ist: jede beste CS pro Ordnung k ist interessant, da sie als Ausgangspunkt für die neue CS der Ordnung $k + 1$ fungiert. Die EpitopeMatch-OF unterscheidet sich von allen anderen bekannten OFs und gestattet eine intuitivere Bewertung der Ähnlichkeit. Der Grund für die allgemeine Verwendung des quadratischen statt des arithmetischen Mittelwerts ist seine Sensitivität gegenüber den Ausreißern. Auch EpitopeMatch basiert auf der RMSD (Gl. 31). Ihr Wertebereich $0\mathring{A} \leq RMSD < \infty\mathring{A}$ wird jedoch zu $0 \leq NRMSD \leq 1$ normalisiert (Gl. 35), sodass die RMSD-Werte durch die sigmoide Verzerrung in drei Zonen eingeteilt werden können: unterhalb, innerhalb und oberhalb der Zwielichtzone (Abb. 10). Die Zoneneinteilung kann über die Verschiebung des Wendepunkts der Kurve getriggert werden. Durch diese Streckung der RMSD-Werte um den Wendepunkt wird eine bessere Vergleichbarkeit der RMSD-Werte erreicht, die sich sowohl auf die Interpretation als auch auf die Performance positiv auswirkt. Die normalisierten Substitutionsgewichte (Gl. 18) werden vor der sigmoiden Normalisierung (Gl. 36) in den RMSD-Ausdruck integriert (Gl. 33). Jede CS der Ordnung $k > 1$ entsteht durch die Erweiterung einer CS der Ordnung $k - 1$ durch eine CS (aus SM) der Ordnung $k = 1$. Diese Verkettung der Substrukturen erlaubt die Implikation der Substrukturinformationen in die Gesamtbewertung der resultierenden CS (Gl. 40), was den Gegenstand der EpitopeMatch-OF ausmacht. Alle geometrieabhängigen EpitopeMatch-Deskriptoren (Abs. 2.2.6.1) basieren auf Normalized Weighted Root Mean Square Deviation (NWRMSD) (Gl. 36). Während die BCS über die Balance zwischen der NWRMSD und der Größe einer CS ermittelt wird (Gl. 37), bewertet der Common Substructure Score (CSS) (Gl. 40) den Inhalt einer gemeinsamen Substruktur deutlich sensitiver als die alleinige RMSD. Jede CS ist im Übrigen eine Clique, die statt in den graphentheoretischen [77], in den neuerdings kombinatorischen Formalismus verpackt ist.

Je kleiner der Ähnlichkeitskern zweier Strukturen ist, desto mehr CSs aus SM sind falsch-positiv. Diese fallen als solche im Verlauf der Kombination der richtig-positiven CSs aus SM zu der BCS automatisch heraus. Die BCS wird anschließend optimiert bzw. erweitert,

indem anhand ihrer Transformationsdaten neue CSs der Ordnung $k = 1$ für die fehlenden Zuordnungen zu den Gruppen der kleineren Struktur ausprobiert werden. Falls die BCS nicht mehr verbessert werden kann, bricht der Algorithmus ab, greift auf den Vorrat der CSs aus SM zurück, die in den bislang ermittelten BCSs nicht enthalten sind und versucht daraus eine neue BCS zu ermitteln. Auf diese Weise werden mehrfach vorhandene Muster oder die einzelnen Domänen mit großen Konformationsänderungen erkannt, die anschließend zu einem Gesamtbild zusammengeführt werden (flexibles Matchen).

Abs. 2.2 enthält eine detaillierte Beschreibung der Heuristik. Abs. 2.3 demonstriert anhand einer Reihe von Beispielen die Qualität und die Bandbreite des Algorithmus. Abs. 2.4 vergleicht die Qualität und die Performance von EpitopeMatch mit den als "State of the Art" geltenden Algorithmen.

2.2 HEURISTIK

EpitopeMatch garantiert weder die optimale Rechenzeit noch die optimale Lösung. Die Rechenzeit hängt stark von der Professionalität der Implementierung und der gewählten Programmiersprache ab. Mit der steigenden Unähnlichkeit der zu vergleichenden Strukturen wächst die Anzahl von Alternativalignments, die sich nur geringfügig voneinander unterscheiden. Die Definition einer einzigen optimalen Lösung ist in diesem Fall nicht sinnvoll. Als ein heuristisches Verfahren geht EpitopeMatch einen Kompromiss zwischen dem Rechenaufwand und der Güte der gefundenen Lösung ein. Um dem einen oder dem anderen Anspruch eher gerecht zu werden, verfügt EpitopeMatch über einen bestimmten Parametersatz, mit dem die Performance und die Qualität gesteuert werden kann. Die folgenden Unterabschnitte beschreiben den Algorithmus.

2.2.1 *Bandbreite*

Der Algorithmus ist im Allgemeinen auf beliebige räumliche Konformationen beliebiger chemischer Verbindungen anwendbar. Er arbeitet auf dem diskreten Datensatz der Strukturinformation, den Atomkoordinaten, und nutzt die Nomenklatur für die templatebasierte Definition der interatomaren Beziehungen. Den aktuellsten Stand der Proteinstrukturaufklärung bietet die PDB. Abgesehen von den Proteinen ist diese Datenbank reich an anderen Biopolymeren und einfachen Molekülen, die als Liganden oder einzeln aufgeklärt sind. Die monomeren chemischen Verbindungen der Liganden werden als eine Teilmenge der PDB in der CCD geführt.

2.2.2 *Kombinatorische Explosion*

Die Ausgangssituation ist das Worst-Case-Szenario, in dem jede Aminosäure der Querystruktur $Q = \{q|q \in Aminosäurenobjekte\}$ jeder Aminosäure der Targetstruktur $T = \{t|t \in Aminosäurenobjekte\}$ entsprechen kann und die Reihenfolge, in der die Aminosäuren der jeweiligen Struktur aneinander gebunden sind, unbekannt ist. Jede Aminosäure ist ein Objekt, das aus mindestens einer Atomkoordinate der entsprechenden Aminosäure besteht. Die Menge

$$S = Q \times T = \{(q,t)\,|q \in Q \wedge t \in T\} \tag{1}$$

enthält alle möglichen Query-Target-Residuenpaare (2er-Tupel) der beiden Strukturen. Jedes 2er-Tupel dieser Menge ist eines der insgesamt

$$|S| = |Q| \cdot |T| \tag{2}$$

Objekte der zweidimensionalen Scoringmatrix (Abs. 2.2.4). Jedes Objekt der Scoringmatrix ist somit ein korrespondierendes Query-Target-Residuenpaar mit der entsprechenden Bewertung der Korrespondenz der beiden Residuen. Mindestens eine Kombination der Objekte der Scoringmatrix ist das gesuchte Alignment. Abb. 3 veranschaulicht die Anzahl aller

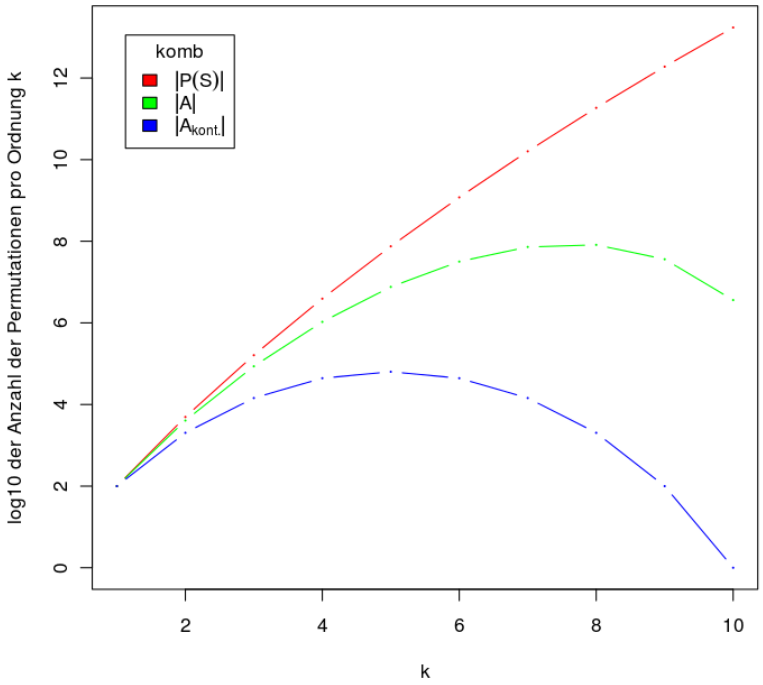


Abbildung 3: Mit $|Q| = |T| = 10$ und $1 \leq k \leq 10$ ist $|P(S)| = 1.941591 \cdot 10^{13}$, $|A| = 2.3466223 \cdot 10^{8}$ und $|A_{kont.}| = 1.84755 \cdot 10^{5}$. $|P(S)|$ in rot, $|A|$ in grün und $|A_{kont.}|$ in blau.

möglichen Kombinationen (rot), die Anzahl aller diskontinuierlichen und kontinuierlichen Kombinationen (grün) und die Anzahl aller kontinuierlichen Kombinationen (blau) zweier Strukturen mit jeweils 10 Aminosäuren. Die Menge aller Kombinationen (Teilmengen) K von S (Abb. 3, rot)

$$P(S) = \{K | K \subset S\} \tag{3}$$

der Ordnungen

$$1 \leq k \leq min(|Q|, |T|)$$

besteht aus

$$|P(S)| = \sum_{k=1}^{min(|Q|,|T|)} \binom{|S|}{k} \tag{4}$$

Kombinationen. Die Variationen der Elemente der jeweiligen Kombination sind in der Menge $P(S)$ nicht enthalten. Die Reihenfolge der korrespondierenden Residuenpaare eines Alignments spielt für seine Bewertung keine Rolle, da ihre Permutation keine Auswirkung auf die Berechnung der RMSD (Gl. 31) hat. Die Menge

$$A = \{K^A | K^A \in P(S)\} \tag{5}$$

aller diskontinuierlichen und kontinuierlichen Alignments (Abb. 3, grün) der beiden Strukturen ist die Menge aller 2er-Tupel-Kombinationen

$$K^A = \{(q,t)|(q,t) \in S \wedge \forall q \, und \, \forall t \, sind \, verschieden\} \tag{6}$$

mit voneinander verschiedenen Query- und voneinander verschiedenen Target-Objekten. Es existieren genau

$$|A| = \sum_{k=1}^{min(|Q|,|T|)} \binom{|Q|}{k} \binom{|T|}{k} k! \tag{7}$$

Alignmentkombinationen der Ordnungen

$$1 \leq k \leq min(|Q|,|T|)$$

Somit ist $|A|$ die maximale Anzahl der zu untersuchenden Alignmentkombinationen und gleichzeitig die maximale Anzahl der möglichen gemeinsamen Substrukturen. Mit den wachsenden Größen der QS und der TS wird sie sehr viel kleiner

$$|A| \lll |P(S)| \tag{8}$$

als die Anzahl der Kombinationen $|P(S)|$ entsprechender Ordnungen k (Abb. 3, grün, rot). Dennoch ist die Explosion der Anzahl ihrer Elemente schon bei kleinen Strukturgrößen deutlich spürbar (Abb. 3, grün).

Sortiert man die Tupel einer beliebigen Alignmentkombination nach q oder t, so lässt sie sich in eine der vier Alignment-Klassen einordnen (Tab. 1). Diese Klassifizierung suggeriert,

K^A-Klasse	Query-Objekte	Target-Objekte	Alignment-Art
1	kont., ggf. mit Gaps	kont., ggf. mit Gaps	kont.
2	kont., ggf. mit Gaps	diskont.	diskont.
3	diskont.	kont., ggf. mit Gaps	diskont.
4	diskont.	diskont.	diskont.

Tabelle 1: Alignment-Klassen

dass die Menge

$$A^{kont.} = \{K^{A,kont.}|K^{A,kont.} \in A\} \tag{9}$$

aller kontinuierlichen Alignments (Abb. 3, blau)

$$K^{A,kont.} = \{(q,t)|(q,t) \in S \wedge \forall q \, und \, \forall t \, sind \, verschieden \, und \, konsekutiv\} \tag{10}$$

mit der Anzahl der Elemente

$$|A^{kont.}| = \sum_{k=1}^{min(|Q|,|T|)} \binom{|Q|}{k} \binom{|T|}{k} \tag{11}$$

mit den wachsenden Strukturgrößen sehr viel kleiner ist, als die Menge aller diskontinuierlichen Alignments (Abb. 3).

$$|A^{kont.}| \lll |A^{diskont.}| \, mit \, |A^{diskont.}| = |A| - |A^{kont.}| \tag{12}$$

Der Unterschied pro Anzahl der Kombinationen aus $A^{kont.}$ und $A^{diskont.}$ der Ordnung k ist der Faktor $k!$.

An dieser Stelle wird die Bedeutung der kombinatorischen Explosion bezüglich der Alignment-Bildung deutlich. Jede Alignmentkombination aus A ist eine potentielle Lösung. Somit ist das Problem NP-hart [94]. Das Generieren aller Alignmentkombinationen aus A im Rahmen einer sequentiellen Abarbeitung (eine CPU) für große Q und T ist schlicht unmöglich. Man ist darauf angewiesen das Worst-Case-Szenario um zusätzliche Informationen zu ergänzen. Jede Berücksichtigung von zusätzlichen Informationen wie die Primär-, Sekundärstruktur, Strukturfragmentierung usw., führt zu einer drastischen Verkleinerung des gesamten Kombinationsraums A und schließt zwangsläufig eine entsprechende Teilmenge von A aus. So impliziert z.B. die Suche nach den kontinuierlichen Alignments aus $A^{kont.}$ die Möglichkeit einer gerichteten Bewegung durch die Scoringmatrix. Eine Reihe von etablierten Algorithmen (Abs. 1.2.1) nutzen diesen Vorteil aus, indem sie die Suche auf Grundlage der kontinuierlichen Fragmente aufbauen. Die strikte Kombination der kontinuierlichen Fragmentpaare resultiert jedoch in den kontinuierlichen gemeinsamen Substrukturen aus dem Kombinationsraum $A^{kont.}$. Für die Suche nach den diskontinuierlichen Alignments aus $A^{diskont.}$ ist die kontinuierliche Fragmentierung nicht geeignet.

EpitopeMatch ist ein heuristisches Verfahren, das im Rahmen einer Reihe von Parametern eine Bewegung im gesamten Kombinationsraum A erlaubt und somit Alignments sowohl aus $A^{kont.}$ als auch aus $A^{diskont.}$ zulässt.

2.2.3 *Distanzmatrizen*

Jede Struktur aus der PDB ist ein statischer Schnappschuss ihres natürlichen dynamischen Verhaltens. EpitopeMatch berücksichtigt keine dynamischen Strukturänderungen, sondern

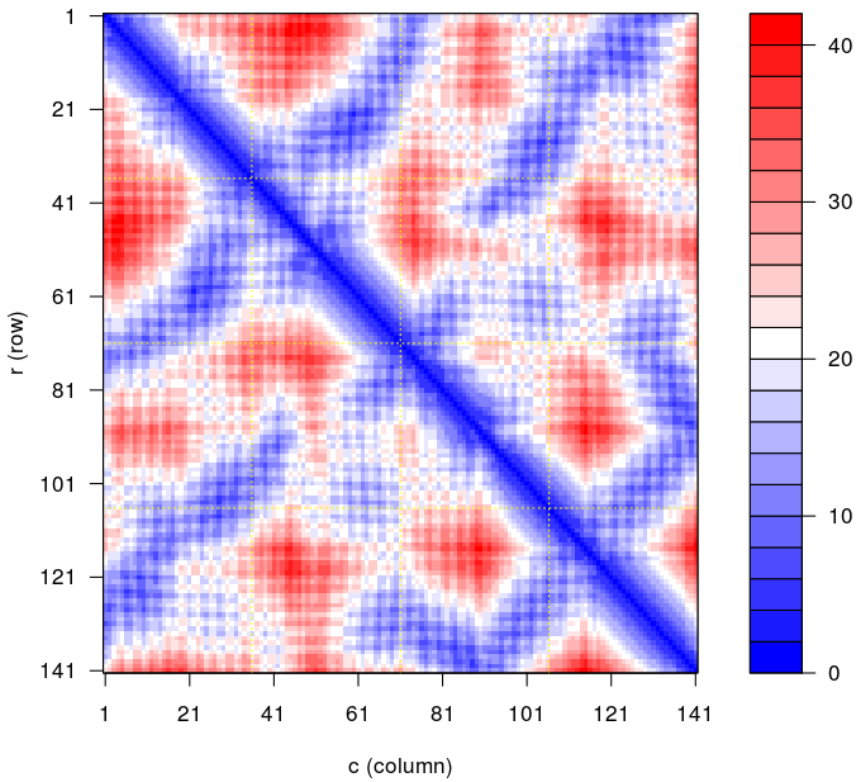


Abbildung 4: Distanzmatrix der Cα-Atome der Kette 2DN2.A. Die Achsen enthalten die Sequenznummern der Residuen. Das Farbspektrum repräsentiert die Distanzen in $\AA$.

vergleicht lediglich die statischen Konformationen der Atome. An diesem Punkt stellt sich

die Frage, welche interatomaren Beziehungen sollen für den Vergleich zweier Strukturen berücksichtigt werden? Dieses Kapitel behandelt explizit die Monomere der Proteine, die proteinogenen Aminosäuren. Jede Aminosäure einer Proteinstruktur besteht aus mindestens vier Rückgratatomen N, Cα, C und O. Holm and Sander [65] repräsentierten die dreidimensionale Struktur eines Proteins in der Form einer zweidimensionalen Distanzmatrix, in der alle Distanzen zwischen den Cα-Atomen der Residuen gespeichert sind. Die Distanzmatrix DM^Q eines Proteins Q mit $|Q|$ Residuen ist eine $|Q| \times |Q|$ Matrix (Abb. 4). Eine Distanz zwischen zwei Cα-Atomen r und c der Struktur Q ist d^Q_{rc}, wobei

$$\begin{array}{ll} 1 \leq r, c \leq |Q| & r \text{ der Reihen-, } c \text{ der Spaltenindex ist,} \\ d^Q_{rc} = d^Q_{cr} & \text{die Distanzmatrix symmetrisch ist,} \\ d^Q_{rc} = 0 & \text{wenn } r = c \text{ ist.} \end{array}$$

Die Distanzmatrixdarstellung hat einen entscheidenden Vorteil, sie ist rotations- und translationsinvariant. Der Vergleich zweier Strukturen anhand ihrer Distanzmatrizen wird allein über die interatomaren Distanzbeziehungen durchgeführt, ohne jeglicher rechenintensiven Strukturüberlagerungen. Ihr Nachteil ist jedoch, sie bildet die Beziehungen zwischen den Residuen anhand eines einzigen Atoms ab. Dabei geht die Information über die räumliche Ausrichtung der Residuen verloren. Ohne dieser Information ist die Suche nach den kleineren Strukturen auf den großen Strukturen, d.h. der Vergleich sich in ihrer Größe stark unterscheidenden Distanzmatrizen, aufgrund des allgemeinen Distanzenrauschens nicht möglich. Ein Beispiel: Man stelle sich zwei Residuen im Raum vor. Ihre Cα-Atome sind zwei Fixpunkte, die ihre Position relativ zueinander nicht verändern. Nun rotiert man die beiden Residuen zufällig um ihre Cα-Atome. Obwohl sich die relative Ausrichtung der beiden Residuen ändert, bleibt diese Tatsache in der Cα-Distanzmatrix unsichtbar. Eine weiterer Nachteil einer Distanzmatrix ist die fehlende Information über die Chiralität der Residuen.

Gemäß Nomenklatur nach IUPAC-IUB [76] verteilen sich die Atome der Seitenketten über die sechs Ebenen β, γ, δ, ε, ζ und η. Mit den vier Atomen des Rückgrats ergeben sich insgesamt zehn Ebenen (Abb. 5). Diese systematische Einteilung der Aminosäurenatome erlaubt eine einfache Definition einer dreidimensionalen Distanzmatrix. Die EpitopeMatch-Distanzmatrix $EMDM^Q$ eines Proteins Q mit $|Q|$ Residuen ist eine $|Q| \times |Q| \times Z$ Matrix. Eine Distanz zwischen zwei Atomen der Struktur Q ist d^Q_{rcz}, wobei

$$\begin{array}{ll} 1 \leq r, c \leq |Q| & r \text{ der Reihen-, } c \text{ der Spaltenindex ist,} \\ z \geq 1 & z \text{ Index der Atomebene ist,} \\ d^Q_{rcz} = d^Q_{crz} & \text{die Distanzmatrix symmetrisch ist,} \\ d^Q_{rcz} = 0 & \text{wenn } r = c \text{ ist.} \end{array}$$

Einige Residuen enthalten mehr als ein Atom in einer bestimmten Ebene (Abb. 5, NH1 und NH2 des ARG, etc.). In diesem Fall werden ihre Koordinaten zu einem geometrischen Zentrum zusammengefasst. Ein Residuum R^Q_q eines Proteins Q enthält abhängig von dem gewählten Template genau $Z(R^Q_q)$ Koordinaten. Ein Residuentupel RT^Q_{rc} eines Proteins Q wird also durch maximal

$$d_{COUNT}(RT^Q_{rc}) = \min(Z(R^Q_r), Z(R^Q_c))$$

Distanzen zwischen den Atomen entsprechender Ebenen repräsentiert. So können z.B. alle GLY-ARG-Paare durch maximal vier

$$d_{COUNT}(RT^Q_{GLY,ARG}) = 4$$

Distanzen und alle TRP-ARG-Paare durch maximal zehn

$$d_{COUNT}(RT^{Q}_{TRP,ARG}) = 10$$

Distanzen repräsentiert werden. Bei dieser Art der Zuordnung der Korrespondenzen han-

components

20 / 20 / 14070

ALA
ARG
ASN
ASP
CYS
GLN
GLU
GLY
HIS
ILE
LEU
LYS
MET
PHE
PRO
SER
THR
TRP
TYR
VAL

ALA|ARG|ASN|ASP|CYS

components and their atoms

3 / 20 4 / 387 20 / 20 167 / 387

20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ALA 1	N	CA	C	O	CB	OXT	H	H2	HA	HB1	HB2	HB3	HXT		
ALA 1	1	2	3	4	5										
ARG 1	N	CA	C	O	CB	CG	CD	NE	CZ	NH1	NH2	OXT	H	H2	HA
ARG 1	1	2	3	4	5	6	7	8	9	10	10				
ASN 1	N	CA	C	O	CB	CG	OD1	ND2	OXT	H	H2	HA	HB2	HB3	HD21
ASN 1	1	2	3	4	5	6	7	7							
ASP 1	N	CA	C	O	CB	CG	OD1	OD2	OXT	H	H2	HA	HB2	HB3	HD2
ASP 1	1	2	3	4	5	6	7	7							
CYS 1	N	CA	C	O	CB	SG	OXT	H	H2	HA	HB2	HB3	HG	HXT	
CYS 1	1	2	3	4	5	6									
GLN 1	N	CA	C	O	CB	CG	CD	OE1	NE2	OXT	H	H2	HA	HB2	HB3
GLN 1	1	2	3	4	5	6	7	8	8						
GLU 1	N	CA	C	O	CB	CG	CD	OE1	OE2	OXT	H	H2	HA	HB2	HB3
GLU 1	1	2	3	4	5	6	7	8	8						
GLY 1	N	CA	C	O	OXT	H	H2	HA2	HA3	HXT					
GLY 1	1	2	3	4											
HIS 1	N	CA	C	O	CB	CG	ND1	CD2	CE1	NE2	OXT	H	H2	HA	HB2
HIS 1	1	2	3	4	5	6	7	7	8	8					
ILE 1	N	CA	C	O	CB	CG1	CG2	CD1	OXT	H	H2	HA	HB	HG12	HG13
ILE 1	1	2	3	4	5	6	6	7							
LEU 1	N	CA	C	O	CB	CG	CD1	CD2	OXT	H	H2	HA	HB2	HB3	HG
LEU 1	1	2	3	4	5	6	7	7							
LYS 1	N	CA	C	O	CB	CG	CD	CE	NZ	OXT	H	H2	HA	HB2	HB3
LYS 1	1	2	3	4	5	6	7	8	9						
MET 1	N	CA	C	O	CB	CG	SD	CE	OXT	H	H2	HA	HB2	HB3	HG2
MET 1	1	2	3	4	5	6	7	8							
PHE 1	N	CA	C	O	CB	CG	CD1	CD2	CE1	CE2	CZ	OXT	H	H2	HA
PHE 1	1	2	3	4	5	6	7	7	8	8	9				
PRO 1	N	CA	C	O	CB	CG	CD	OXT	H	HA	HB2	HB3	HG2	HG3	HD2
PRO 1	1	2	3	4	5	6	7								
SER 1	N	CA	C	O	CB	OG	OXT	H	H2	HA	HB2	HB3	HG	HXT	
SER 1	1	2	3	4	5	6									
THR 1	N	CA	C	O	CB	OG1	CG2	OXT	H	H2	HA	HB	HG1	HG21	HG22
THR 1	1	2	3	4	5	6	6								
TRP 1	N	CA	C	O	CB	CG	CD1	CD2	NE1	CE2	CE3	CZ2	CZ3	CH2	OXT
TRP 1	1	2	3	4	5	6	7	7	8	8	8	9	9	10	
TYR 1	N	CA	C	O	CB	CG	CD1	CD2	CE1	CE2	CZ	OH	OXT	H	H2
TYR 1	1	2	3	4	5	6	7	7	8	8	9	10			
VAL 1	N	CA	C	O	CB	CG1	CG2	OXT	H	H2	HA	HB	HG11	HG12	HG13
VAL 1	1	2	3	4	5	6	6								

Abbildung 5: AllAtoms-Template der 20 proteinogenen Aminosäuren (sortiert nach dem Dreibuchstabencode). Die Residuenatome verteilen sich auf bis zu 10 Ebenen: N, Cα, C und O (Rückgratatome); β, γ, δ, ε, ζ und η (Restatome). Jede Ebene erhält eine Nummer (1 - 10). Jedes Nicht-Wasserstoff-Atom erhält die Nummer der entsprechenden Ebene. Mehr als ein Atom in einer Ebene (z.B. ARG Nη1 & Nη1) bedeutet ihre Zusammenfassung zu einem geometrischen Zentrum. Atome einer Ebene gelten als korrespondierend und werden von einer Distanzmatrix repräsentiert. AllAtoms-Template wird von insgesamt 10 Distanzmatrizen repräsentiert. Die Korrespondenzen sind beliebig wählbar.

delt es sich lediglich um einen Vorschlag. Die templatebasierte Zuordnung der Residuenatome erlaubt die freie Korrespondenzendefinition.

Eine weitere Eigenschaft eines Residuentupels ist der Distanzenmittelwert

$$d_{MEAN}(RT^{Q}_{rc}) = \frac{d_{SUM}(RT^{Q}_{rc})}{d_{COUNT}(RT^{Q}_{rc})} \tag{13}$$

Da die Distanzmatrix symmetrisch ist, enthält ein Protein Q genau $\frac{|Q|^2-|Q|}{2}$ unterschiedliche Residuentupel.

2.2.4 *Scoringmatrix*

Existiert eine zweite Struktur T, mit der EpitopeMatch-Distanzmatrix $EMDM^T$, so kann eine Scoringmatrix SM^{QT} mit $|Q| \times |T|$ Objekten definiert werden. Jedes Objekt dieser Matrix ist ein Residuentupel $RT_{qt}^{QT}(R_q^Q, R_t^T)$. Jedes Residuentupel besitzt die folgenden Eigenschaften:

$0 \leq COMPL(RT_{qt}^{QT}) \leq 1$ completeness (COMPL)

$0 \leq SSIM(RT_{qt}^{QT}) \leq 1$ Substitution SIMilarty (SSIM)

$0 \leq RMSS(RT_{qt}^{QT}) \leq 1$ Root Mean Square Similarity (RMSS)

Sollen ein Query-Residuum R_q^Q und ein Target-Residuum R_t^T einander entsprechen, so kann man erwarten, dass sie in der Anzahl ihrer Atome bzw. Koordinaten möglichst gleich sind. Je mehr sie sich in der Anzahl ihrer Koordinaten unterscheiden, desto geringer ist die Wahrscheinlichkeit, dass sie einander entsprechen. Dieser, sich an die Geometrie und Größe der Residuen anlehnende Parameter ist definiert als

$$COMPL(RT_{qt}^{QT}) = \frac{\min(Z(R_q^Q), Z(R_t^T))}{\max(Z(R_q^Q), Z(R_t^T))} \tag{14}$$

So ist z.B. die Komponentenvollständigkeit eines GLY-ARG-Tupels

$$COMPL(RT_{GLY,ARG}^{QT}) = \frac{4}{10} = 0.4$$

EpitopeMatch sieht die Möglichkeit vor, ein Residuentupel zusätzlich zu gewichten. Die

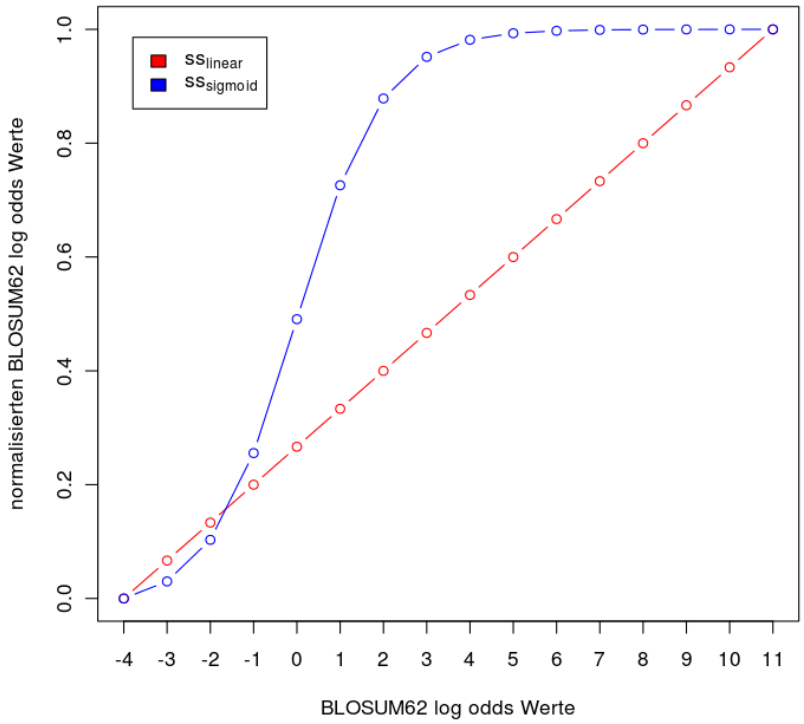


Abbildung 6: Normalisierung der Blosum62-"log-odds"-Werte linear (rot) und sigmoid (blau).

Definition der Substitutionsähnlichkeit eines Residuentupels liegt im Ermessen des Anwenders. Bei den Aminosäuren der Proteine greift EpitopeMatch auf die auf empirischen Daten beruhenden Substitutionsmatrizen von Dayhoff et al. [42] und Henikoff and Henikoff [61] zurück. Sie geben pro Aminosäuren-Paar eine Substitutionswahrscheinlichkeit an

und unterscheiden sich in der Betrachtung der relativen Mutationsraten verwandter Proteine. Die Substitutionswahrscheinlichkeit eines Aminosäuren-Paares wird im EpitopeMatch-Kontext als seine physiko-chemische Substitutionsähnlichkeit interpretiert. Am häufigsten verwendet ist die BLOSUM62-Matrix BM^{62}. Die "log-odds"-Werte dieser Matrix sind für den additiven Gebrauch konzipiert und schwanken zwischen -4 und 11. Für ihre multiplikative Verwendung durch EpitopeMatch müssen sie normalisiert werden. Eine direkte Interpretation der linear normalisierten BM^{62}-Werte (Abb. 6 rot)

$$SSIM_{linear}(RT_{qt}^{QT}) = \frac{BM^{62}(RT_{qt}^{QT}) - \min(BM^{62})}{\max(BM^{62}) - \min(BM^{62})} \tag{15}$$

als physiko-chemische Ähnlichkeit zweier Residuen ist nicht möglich. Ihre Schwankung zwischen $0.5\overline{3}$ und 1.0 auf der Hauptdiagonale der so normalisierten Matrix führt zu der Annahme, dass die physiko-chemische Ähnlichkeit z.B. eines ALA-ALA-Tupels mit

$$SSIM_{linear}(RT_{ALA,ALA}^{QT}) = 0.5\overline{3}$$

lediglich $53.\overline{3}\%$, statt der erwarteten 100% ist. Aus diesem Grund werden die BM^{62}-Werte sigmoid exponenziert

$$BM_{sigmoid}^{62} = \frac{1}{1 + e^{-BM^{62}}} \tag{16}$$

sodass die daraus resultierende Substitutionsähnlichkeit (Abb. 6 blau)

$$SSIM(RT_{qt}^{QT}) = \frac{BM_{sigmoid}^{62}(RT_{qt}^{QT}) - \min(BM_{sigmoid}^{62})}{\max(BM_{sigmoid}^{62}) - \min(BM_{sigmoid}^{62})} \tag{17}$$

auf der Hauptdiagonale ≈ 1.0 bzw. $\approx 100\%$ ist (Abb. 7).

20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	ALA 1	ARG 1	ASN 1	ASP 1	CYS 1	GLN 1	GLU 1	GLY 1	HIS 1	ILE 1	LEU 1	LYS 1	MET 1	PHE 1	PRO 1	SER 1	THR 1	TRP 1	TYR 1	VAL 1
ALA 1	0.982	0.256	0.103	0.103	0.491	0.256	0.256	0.491	0.103	0.256	0.256	0.256	0.256	0.103	0.256	0.726	0.491	0.030	0.103	0.491
ARG 1	0.256	0.993	0.491	0.103	0.030	0.726	0.491	0.103	0.491	0.030	0.103	0.879	0.256	0.030	0.103	0.256	0.256	0.030	0.103	0.030
ASN 1	0.103	0.491	0.997	0.726	0.030	0.491	0.491	0.491	0.726	0.030	0.030	0.491	0.103	0.030	0.103	0.726	0.491	0.000	0.103	0.030
ASP 1	0.103	0.103	0.726	0.997	0.030	0.491	0.879	0.256	0.256	0.030	0.000	0.256	0.030	0.030	0.256	0.491	0.256	0.000	0.030	0.030
CYS 1	0.491	0.030	0.030	0.030	1.000	0.030	0.000	0.030	0.030	0.256	0.256	0.030	0.256	0.103	0.030	0.256	0.256	0.103	0.103	0.256
GLN 1	0.256	0.726	0.491	0.491	0.030	0.993	0.879	0.103	0.491	0.030	0.103	0.726	0.491	0.030	0.256	0.491	0.256	0.103	0.256	0.103
GLU 1	0.256	0.491	0.491	0.879	0.000	0.879	0.993	0.103	0.491	0.030	0.030	0.726	0.103	0.030	0.256	0.491	0.256	0.030	0.103	0.103
GLY 1	0.491	0.103	0.491	0.256	0.030	0.103	0.103	0.997	0.103	0.000	0.000	0.103	0.030	0.030	0.103	0.491	0.103	0.103	0.030	0.030
HIS 1	0.103	0.491	0.726	0.256	0.030	0.491	0.491	0.103	1.000	0.030	0.030	0.256	0.103	0.256	0.103	0.256	0.103	0.103	0.879	0.030
ILE 1	0.256	0.030	0.030	0.030	0.256	0.030	0.030	0.000	0.030	0.982	0.879	0.030	0.726	0.491	0.030	0.103	0.256	0.030	0.256	0.952
LEU 1	0.256	0.103	0.030	0.000	0.256	0.103	0.030	0.000	0.030	0.879	0.982	0.103	0.879	0.491	0.030	0.103	0.256	0.103	0.256	0.726
LYS 1	0.256	0.879	0.491	0.256	0.030	0.726	0.726	0.103	0.256	0.030	0.103	0.993	0.256	0.030	0.256	0.491	0.256	0.030	0.103	0.103
MET 1	0.256	0.256	0.103	0.030	0.256	0.491	0.103	0.030	0.103	0.726	0.879	0.256	0.993	0.491	0.103	0.256	0.256	0.256	0.256	0.726
PHE 1	0.103	0.030	0.030	0.030	0.103	0.030	0.030	0.030	0.256	0.491	0.491	0.030	0.491	0.997	0.000	0.103	0.103	0.726	0.952	0.256
PRO 1	0.256	0.103	0.103	0.256	0.030	0.256	0.256	0.103	0.103	0.030	0.030	0.256	0.103	0.000	0.999	0.256	0.256	0.000	0.030	0.103
SER 1	0.726	0.256	0.726	0.491	0.256	0.491	0.491	0.491	0.256	0.103	0.103	0.491	0.256	0.103	0.256	0.982	0.726	0.030	0.103	0.103
THR 1	0.491	0.256	0.491	0.256	0.256	0.256	0.256	0.103	0.103	0.256	0.256	0.256	0.256	0.103	0.256	0.726	0.993	0.103	0.103	0.491
TRP 1	0.030	0.030	0.000	0.000	0.103	0.103	0.030	0.103	0.103	0.030	0.103	0.030	0.256	0.726	0.000	0.030	0.103	1.000	0.879	0.030
TYR 1	0.103	0.103	0.103	0.030	0.103	0.256	0.103	0.030	0.879	0.256	0.256	0.103	0.256	0.952	0.030	0.103	0.103	0.879	0.999	0.256
VAL 1	0.491	0.030	0.030	0.030	0.256	0.103	0.103	0.030	0.030	0.952	0.726	0.103	0.726	0.256	0.103	0.103	0.491	0.030	0.256	0.982

Abbildung 7: Blosum62-Sigmoid Substitutionsmatrix.

Nun lassen sich die Komponentenvollständigkeit und Substitutionsähnlichkeit zu dem Gewicht eines Residuentupels wie folgt zusammenfassen

$$RMSS(RT_{qt}^{QT}) = \begin{cases} \sqrt{\frac{COMPL(RT_{qt}^{QT})^2 + SSIM(RT_{qt}^{QT})^2}{2}} & \text{wenn SSIM größenunabhängig} \\ SSIM(RT_{qt}^{QT}) & \text{wenn SSIM größenabhängig} \\ COMPL(RT_{qt}^{QT}) & \text{sonst} \end{cases} \tag{18}$$

Die Komponentenvollständigkeit und Substitutionsähnlichkeit werden im gewichteten Fall als gleichwertige Eigenschaften betrachtet, wenn die Substitutionsähnlichkeit von der Größe der Residuen unabhängig ist. Ist einer der beiden Werte gleich 1.0, so ist das Tupelgewicht mindestens $\sqrt{2}/2 = 0.7071068$. Dies hebt die Residuentupel mit der gleichen Anzahl der Atome und/oder den gleichen physiko-chemischen Eigenschaften, im Vergleich zu den restlichen Wertekombinationen der COMPL und SSIM besonders hervor.

2.2.5 *Initiales Alignment*

Existiert eine bemerkenswerte Ähnlichkeit zwischen zwei Strukturen Q und T, dann kann diese durch mindestens eine Alignmentkombination K^A (siehe 2.2.2) aus maximal $min(|Q|, |T|)$ Residuentupel RT_{qt}^{QT} ausgedrückt werden. Wenn die Strukturen Q und T identisch sind, dann existiert eine einzige, optimale Alignmentkombination. Wenn die Strukturen Q und T nicht identisch, jedoch einander ähnlich sind, dann existieren oft mehrere alternativen Alignmentkombinationen, die als Ausdruck der Ähnlichkeit in Frage kommen können. Die alternativen Alignmentkombinationen besitzen meistens einen gemeinsamen Kern und unterscheiden sich in nur wenigen Residuentupeln. Die Aufgabe dieser Algorithmusstufe besteht darin, die vermeintlichen richtig-positiven Residuentupel RT_{qt}^{QT} der alternativen Alignmentkombinationen K^A aus der Menge aller Residuentupel der Scoringmatrix SM^{QT} durch geeignetes Scoring hervorzuheben, um den gesamten Kombinationsraum A (siehe 2.2.2) drastisch zu reduzieren.

2.2.5.1 *Quadrupel-Koeffizient*

Eine Query-Target-Residuenzuordnung geschieht immer anhand eines Query-Residuentupels $RT_{qr,qc}^{Q}$ und eines Target-Residuentupels $RT_{tr,tc}^{T}$ mit ähnlichen interatomaren Distanzen. Die Bewertung eines Residuentupels $RT_{qt}^{QT}(R_q^Q, R_t^T)$ erfolgt also über die Auswertung eines Residuen-Quadrupels $RQ(RT_{qr,qc}^{Q}, RT_{tr,tc}^{T})$ mit qr, qc als Reihen-, Spaltenindex der Matrix $EMDM^Q$ und tr, tc als Reihen-, Spaltenindex der Matrix $EMDM^T$. Aufgrund der symmetrischen Natur der Distanzmatrizen existieren genau

$$\frac{(|Q|^2 - Q) \cdot (|T|^2 - T)}{4} \tag{19}$$

unterschiedliche Residuen-Quadrupel. Lediglich ein Bruchteil aller Residuen-Quadrupel besitzt ähnliche Distanzen zwischen den Atomen der Query- und Target-Residuen. Ein Residuen-Quadrupel ist für die Auswertung interessant, wenn

$$RQ(RT_{qr,qc}^{Q}, RT_{tr,tc}^{T}) = \begin{cases} TRUE & d_{qr,qc,z}^{Q} \leqq dt \wedge d_{tr,tc,z}^{T} \leqq dt \wedge |d_{qr,qc,z}^{Q} - d_{tr,tc,z}^{T}| \leqq ddt \\ FALSE & sonst \end{cases} \tag{20}$$

die Distanzen der Query- und Target-Residuentupel kleiner als Distanz-Threshold $dt = 30.0\mathring{A}$ und die Distanzdifferenzen zwischen den Distanzen der Query- und Target-Residuentupel kleiner als Distanzdifferenz-Threshold $ddt = 1.0\mathring{A}$ sind. Die Threshold-Distanz von $30.0\mathring{A}$ wird auf die typische Domänengröße der globulären Proteine von $\approx$ 150 Aminosäuren [186] zurückgeführt. Je weiter zwei Residuen der gesuchten Struktur voneinander entfernt sind, desto höher ist die Wahrscheinlichkeit, dass ihre relative Positionsänderung (Distanzdifferenz) auf der zu durchsuchenden Struktur so hoch ist, dass sie aus dem allgemeinen Distanzenrauschen nicht mehr herausgefiltert werden kann. Durch die Veränderung

des Distanzdifferenz-Thresholds kann das Induced-Fit [95] (Abs. 2.2.7) direkt angesprochen werden. Der Einsatz der beiden Thresholds verkleinert die Menge der Residuen-Quadrupel auf weniger als 1% und führt zu einer drastischen Reduktion des Distanzenrauschens und einer deutlichen Performance-Steigerung.

Der Koeffizient, mit dem ein Residuen-Quadrupel bewertet werden kann, setzt sich aus drei Komponenten zusammen:

1. Distanzkomponente

$$RQ^{TRUE}_{d_{mean}}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) = \frac{1}{2Z}\sum_{z=1}^{Z} d^{Q}_{qr,qc,z} + d^{T}_{tr,tc,z} \tag{21}$$

2. Distanzdifferenzkomponente

$$RQ^{TRUE}_{dd_{mean}}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) = \frac{1}{Z}\sum_{z=1}^{Z} |d^{Q}_{qr,qc,z} - d^{T}_{tr,tc,z}| \tag{22}$$

3. Vollständigkeitskomponente
   Query- und Target-Residuentupel besitzen gemeinsam

$$\min(d_{COUNT}(RT^{Q}_{qr,qc}), d_{COUNT}(RT^{T}_{tr,tc}))$$

   Paare ähnlicher Distanzen. Unter der Voraussetzung, dass möglichst alle Paare einander entsprechen müssen, wird die Distanzenvollständigkeit wie folgt definiert

$$RQ^{TRUE}_{d_{compl.}}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) = \frac{\min(d_{COUNT}(RT^{Q}_{qr,qc}), d_{COUNT}(RT^{T}_{tr,tc}))}{\max(d_{COUNT}(RT^{Q}_{qr,qc}), d_{COUNT}(RT^{T}_{tr,tc}))} \tag{23}$$

Der Residuen-Quadrupel-Koeffizient wird dann wie folgt berechnet

$$RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) = \frac{dt - RQ^{TRUE}_{d_{mean}}}{dt} \cdot \frac{ddt - RQ^{TRUE}_{dd_{mean}}}{ddt} \cdot RQ^{TRUE}_{d_{compl.}} \tag{24}$$

Je größer die mittlere Entfernung, die mittlere Distanzdifferenz und die Abweichung in der Anzahl der gemeinsamen Distanzen ist, desto kleiner ist der Residuen-Quadrupel-Koeffizient.

2.2.5.2 *Tupel-Score*

Sind die interatomaren Distanzen eines Query-Residuentupels $RT^{Q}_{qr,qc}$ ähnlich den interatomaren Distanzen eines Target-Residuentupels $RT^{T}_{tr,tc}$, d.h, existiert ein Residuen-Quadrupel mit einem großen Koeffizient $RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc})$, so kann man annehmen, dass die Query-Target-Residuentupel

$RT^{QT}_{qr,tr}$ und $RT^{QT}_{qc,tc}$	in paralleler Richtung
$RT^{QT}_{qr,tc}$ und $RT^{QT}_{qc,tr}$	in antiparalleler Richtung
$RT^{QT}_{qr,tr}$, $RT^{QT}_{qc,tc}$, $RT^{QT}_{qr,tc}$ und $RT^{QT}_{qc,tr}$	richtungsinvariant

richtig-positive Tupel sind. Die Zuordnungsrichtung wird im Kontext der Proteine als

$Q(N' \to C') \hat{=} T(N' \to C')$	parallel
$Q(N' \to C') \hat{=} T(C' \to N')$	antiparallel
zum Teil $Q(N' \to C') \hat{=} T(N' \to C')$ und zum Teil $Q(N' \to C') \hat{=} T(C' \to N')$	invariant

zu der Syntheserichtung der Translation an den Ribosomen verstanden, wobei N' der Amino- und C' der Carboxy-Terminus ist. Die Scoringmatrix wird für alle existierenden Quadrupel $RQ^{TRUE}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc})$ iterativ berechnet. Die Scores werden je nach Wahl des Benutzers verteilt:

1. parallel

$$SM^{QT}_{qr,tr} = RMSS(RT^{QT}_{qr,tr}) \cdot \sum RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) \cdot RMSS(RT^{QT}_{qc,tc}) \tag{25}$$

$$SM^{QT}_{qc,tc} = RMSS(RT^{QT}_{qc,tc}) \cdot \sum RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) \cdot RMSS(RT^{QT}_{qr,tr}) \tag{26}$$

2. antiparallel

$$SM^{QT}_{qr,tc} = RMSS(RT^{QT}_{qr,tc}) \cdot \sum RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) \cdot RMSS(RT^{QT}_{qc,tr}) \tag{27}$$

$$SM^{QT}_{qc,tr} = RMSS(RT^{QT}_{qc,tr}) \cdot \sum RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) \cdot RMSS(RT^{QT}_{qr,tc}) \tag{28}$$

3. invariant (parallel & antiparallel)

$$SM^{QT}_{qr,tr} = RMSS(RT^{QT}_{qr,tr}) \cdot \sum RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) \cdot RMSS(RT^{QT}_{qc,tc})$$

$$SM^{QT}_{qc,tc} = RMSS(RT^{QT}_{qc,tc}) \cdot \sum RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) \cdot RMSS(RT^{QT}_{qr,tr})$$

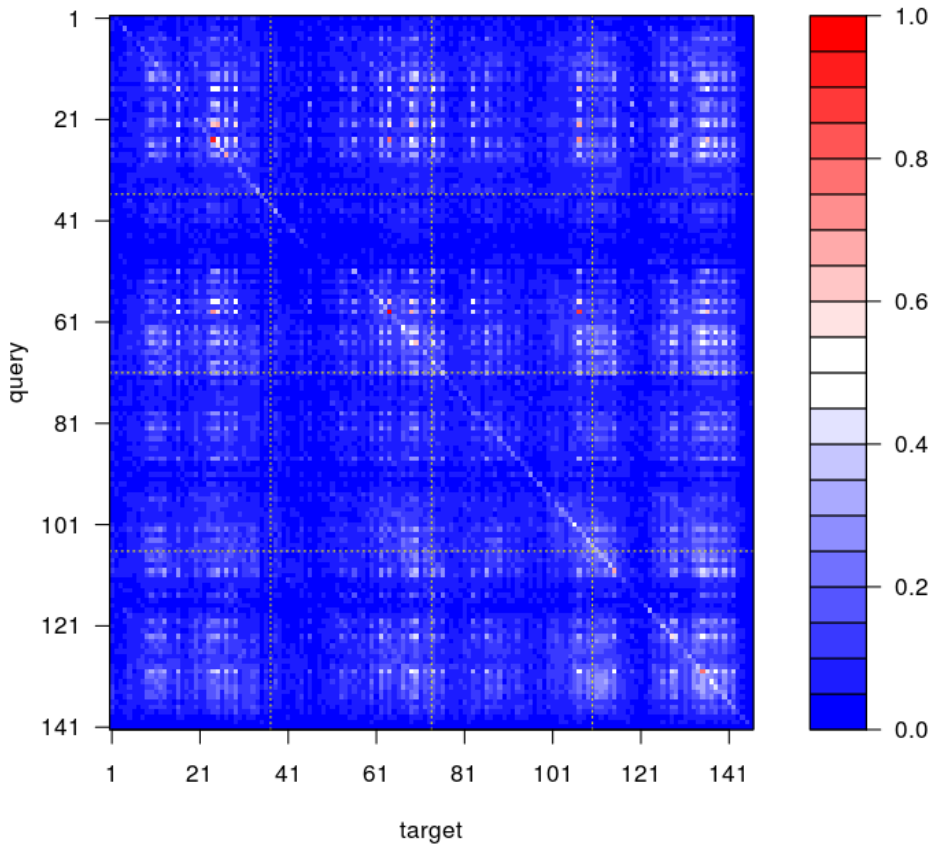
$$SM^{QT}_{qr,tc} = RMSS(RT^{QT}_{qr,tc}) \cdot \sum RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) \cdot RMSS(RT^{QT}_{qc,tr})$$

$$SM^{QT}_{qc,tr} = RMSS(RT^{QT}_{qc,tr}) \cdot \sum RQ^{TRUE}_{coeff.}(RT^{Q}_{qr,qc}, RT^{T}_{tr,tc}) \cdot RMSS(RT^{QT}_{qr,tc})$$

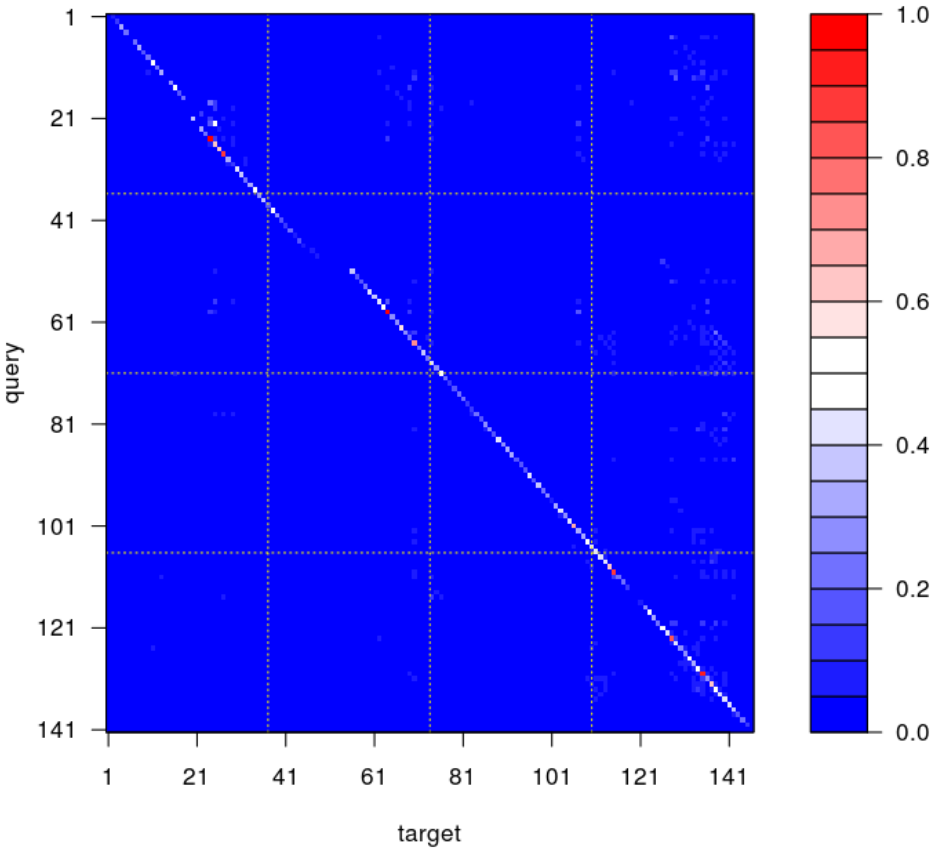
2.2.5.3 *Signalstärke*

Trotz der Einschränkungen durch Thresholds bleibt das Distanzenrauschen allgegenwärtig. Die Abb. 8a zeigt die Scoringmatrix aus dem Vergleich der Kette A mit 141 Aminosäuren und der Kette B mit 146 Aminosäuren aus der Struktur 2DN2. Die Sequenzidentität der beiden Ketten liegt bei 44%. Die Faltungsmuster sind einander sehr ähnlich. Die beiden Strukturen binden den gleichen Liganden und gelten als homolog. Die Distanzmatrix der Kette A ist in der Abb. 4 dargestellt. Die Distanzmatrix der Kette B ist ihr sehr ähnlich. Die Scores SM^{QT}_{qt} der Scoringmatrix in der Abb. 8a sind gegen $max(SM^{QT}_{qt})$ normalisiert. Die Scores der richtig-positiven Residuentupel $RT^{QT,TP}_{qt}$ sind auf der Matrixdiagonale gut erkennbar. Allerdings tauchen sie in der Menge der ähnlich hoch bewerteten falsch-positiven Residuentupel $RT^{QT,FP}_{qt}$ unter, sodass sie nicht eindeutig identifiziert werden können, weil keine Bewegungsrichtung durch die Matrix definiert ist. Die Lösung für dieses Problem ist denkbar einfach - Signale der niedrig bewerteten Residuentupel werden solange aus der Matrix entfernt, bis jedem Query-Residuum nur noch ein Target-Residuum zugeordnet bleibt.

Im Kontext von EpitopeMatch ist die QS immer die kleinere Struktur. In diesem Fall können jedem Query-Residuum maximal $|T|$ Target-Residuen zugeordnet werden. Diese



(a) Verrauschte Scoringmatrix.



(b) Gefilterte Scoring-Summenmatrix.

Abbildung 8: Scoringmatrizen der Ketten 2DN2.A (Query) und 2DN2.B (Target). Die Achsen enthalten die Sequenznummern der Residuen. Das Farbspektrum repräsentiert die normalisierten Scores.

Zuordnung ist durch die jeweilige Reihe der Scoringmatrix repräsentiert. Für jede Matrixreihe wird ein Score-Mittelwert

$$MEAN_q(SM_{qt}^{QT}) = \frac{1}{|T|}\sum_{t=1}^{|T|} SM_{qt}^{QT}, \text{ für alle } SM_{qt}^{QT} > 0 \tag{29}$$

berechnet, sodass

$$RT_{qt}^{QT} = \begin{cases} TP & SM_{qt}^{QT} \geqq MEAN_q(SM_{qt}^{QT}) \\ FP & SM_{qt}^{QT} < MEAN_q(SM_{qt}^{QT}) \end{cases} \tag{30}$$

alle Residuentupel einer Reihe, deren Score kleiner als der Score-Mittelwert ist, aus der weiteren Berechnung ausgeschlossen werden. Die Scores, die in der ersten Iteration berechnet worden sind, werden in der Scoring-Summenmatrix $SSM_{qt}^{QT} = SSM_{qt}^{QT} + SM_{qt}^{QT}$ gespeichert. In der nächsten Iteration wird die Scoringmatrix SM_{qt}^{QT} , unter Ausschluss aller als falsch-positiv markierten Residuentupel $RT_{qt}^{QT,FP}$, erneut berechnet und auf die Scoring-Summenmatrix addiert. Der iterative Prozess bricht ab, sobald in jeder Reihe der Scoringmatrix nur noch ein Residuentupel als richtig-positiv markiert ist. Die Abb. 8b zeigt die resultierende Scoring-Summenmatrix. Das Alignment der beiden Ketten besteht aus

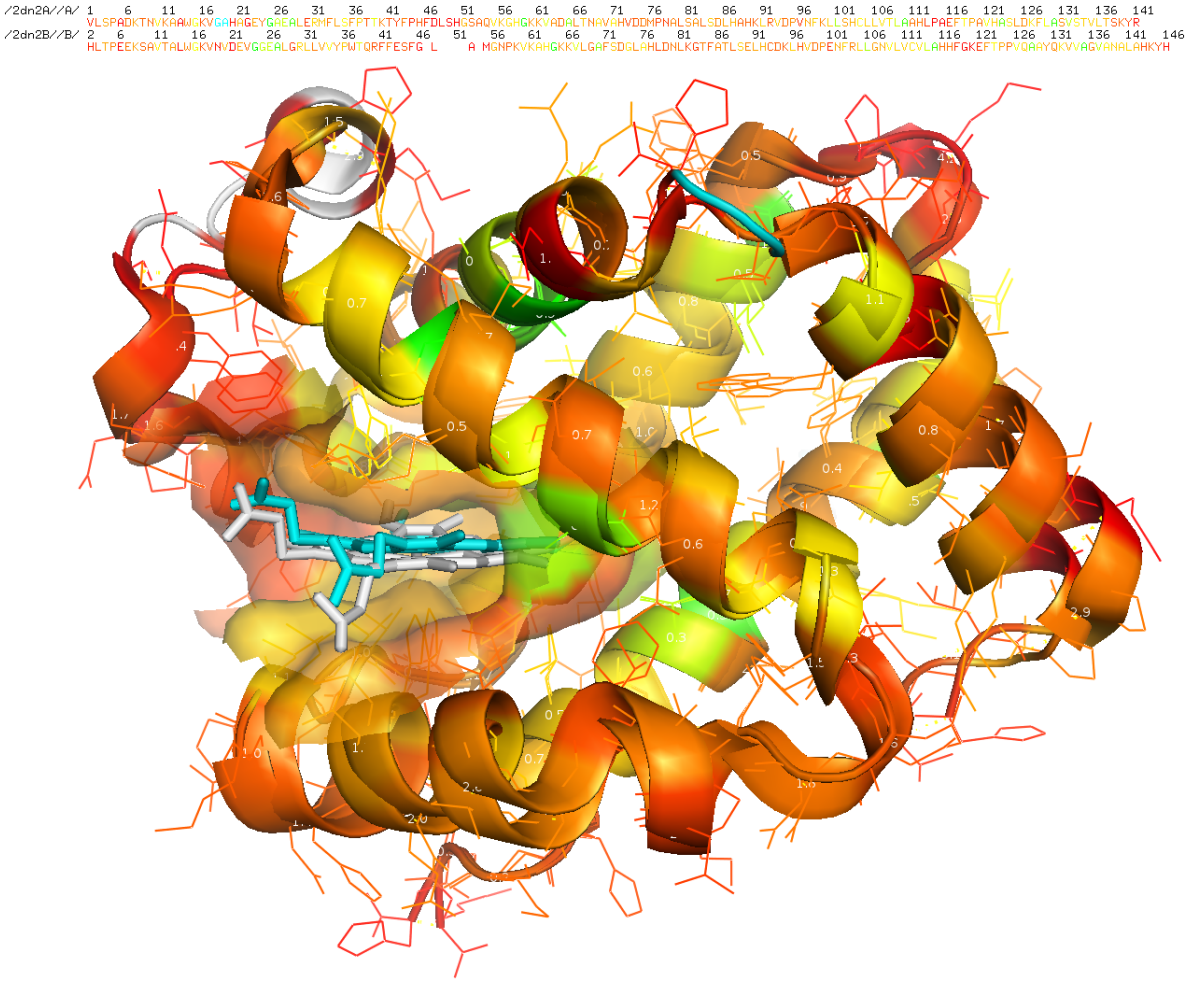


Abbildung 9: Scores der richtig-positiven Residuen. Die am besten bewerteten Residuentupel (min max) befinden sich in den am höchsten konservierten Strukturbereichen.

139 richtig-positiven Residuentupel. Im Fall “parallel” werden 137 und im Fall “invariant”

133 richtig-positive Residuentupel erkannt. Diese Residuentupel werden in der nächsten Algorithmusstufe zum vollständigen Alignment ausgebaut. Überträgt man die Scores der richtig-positiven Tupel auf die überlagerten Strukturen (Abb. 9), dann sieht man erwartungsgemäß, dass die Residuentupel im Inneren der Strukturen höher bewertet sind. Dafür ist die höhere Dichte ihrer Nachbarn verantwortlich.

Die Rechenzeit beträgt im Fall "parallel" $\approx 350ms$ und im Fall "invariant" $\approx 500ms$. Je ähnlicher die beiden Strukturen einander sind, desto schneller werden ihre Gemeinsamkeiten erkannt. Im allgemeinen verfügt dieser Ansatz über die folgenden Eigenschaften:

1. Der Algorithmus terminiert immer.
2. Anwendbar auf beliebige Biopolymere: verzweigt oder unverzweigt; und ihre Substrukturen: kontinuierlich oder diskontinuierlich.
3. Die Erkennung von richtig-positiven Zuordnungen allein anhand von Distanzdifferenzen.
4. Behandlung von Strukturen der typischen Domänengröße von $\approx$ 150 Aminosäuren innerhalb von $\approx 500ms$.
5. Die Scoring-Summenmatrix enthält unter Umständen Informationen von mehreren Alignments (mehrere Epitope).

2.2.6 *Kombinatorisches Resampling*

Die erste Algorithmusstufe liefert eine Alignmentkombination $K^A \in A$ mit $min(|Q|, |T|)$ Residuentupel RT^{QT}. Sie garantiert nicht, dass alle Residuentupel richtig-positiv sind. In den meisten Fällen ist die Menge der richtig-positiven Residuentupel jedoch so groß, dass sie zumindest den Ähnlichkeitskern der beiden Strukturen abdeckt. Die Aufgaben der zweiten Algorithmusstufe sind: die Suche nach den Strukturalignments auf der Grundlage der initialen Alignmentkombination; und die Optimierung der Strukturalignments unter Einsatz von geeigneten Deskriptoren. Die Deskriptoren sind Ähnlichkeitskoeffizienten, die unter dem Aspekt der Plausibilität, der Simplizität und der Wiederverwendbarkeit entwickelt worden sind.

2.2.6.1 *Deskriptoren*

Eine Alignmentkombination $K^A \in A$ der Residuentupel RT^{QT} ist eine bijektive Abbildung von $Q^A \subseteq Q$ nach $T^A \subseteq T$. Die beiden Mengen Q^A und T^A sind gleichmächtig ($|Q^A| = |T^A|$). Unter dieser Voraussetzung, und unter Einsatz der Singulärwertzerlegung [84, 133], werden Transformationsdaten in Form einer Rotationsmatrix R und eines Translationsvektors $\vec{T}$ bestimmt. Diese Transformationsdaten garantieren die optimale Überlagerung der beiden Strukturen bzw. die minimale RMSD, die nach der Transformation der QS

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((R \cdot \vec{q}_i + \vec{T}) - \vec{t}_i)^2} \tag{31}$$

berechnet werden kann und Auskunft über die mittlere Distanzabweichung zwischen den alignierten Atomen der beiden überlagerten Strukturen liefert. Aufgrund der Verteilung der

zu transformierenden Koordinaten auf die Residuenobjekte wird dieser Ausdruck wie folgt modifiziert

$$RMSD(K^A) = \sqrt{\frac{1}{\sum_{k=1}^{|K^A|} Z(RT_k^{QT})} \sum_{k=1}^{|K^A|} \sum_{z=1}^{Z(RT_k^{QT})} ((R \cdot \vec{q_{kz}} + \vec{T}) - \vec{t_{kz}})^2} \tag{32}$$

mit

$1 \leq k \leq |K^A|$ Index der Residuentupel
$1 \leq Z(RT_k^{QT})$ Anzahl der korrespondierenden Koordinaten eines Residuentupels
$1 \leq z \leq Z(RT_k^{QT})$ Index der Koordinaten eines Residuentupels

Unter der Berücksichtigung des Tupelgewichts wird die gewichtete RMSD als

$$WRMSD(K^A) = \sqrt{\frac{1}{\sum_{k=1}^{|K^A|} Z(RT_k^{QT})} \sum_{k=1}^{|K^A|} \frac{1}{RMSS(RT_k^{QT})} \sum_{z=1}^{Z(RT_k^{QT})} ((R \cdot \vec{q_{kz}} + \vec{T}) - \vec{t_{kz}})^2} \tag{33}$$

definiert, mit

$0 < RMSS(RT_k^{QT}) \leq 1$ Tupelgewicht k-tes Residuentupels

Je größer die Distanzabweichungen und je kleiner die Tupelgewichte, desto größer die Weighted Root Mean Square Deviation (WRMSD).

Allgemein stellt sich die Frage, welche Bedeutung haben die RMSD-Werte der überlagerten

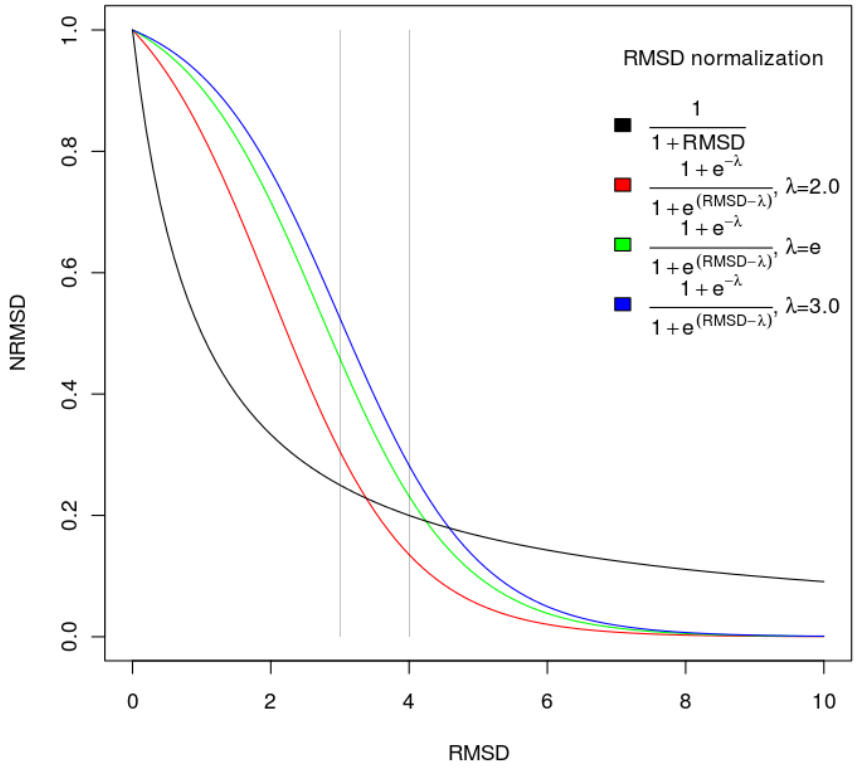


Abbildung 10: Normalisierte RMSD. Die grauen vertikalen Linien markieren die Zwielichtzone. Die Häufigkeit der Alignments innerhalb und oberhalb dieser Zone nimmt drastisch zu.

Strukturen? RMSD-Werte erstrecken sich im rechtsoffenen Intervall $0Å \leq RMSD < \infty Å$.

Erfahrungsgemäß befinden sich die Strukturen mit $0Å \leq RMSD \leq 3Å$ unterhalb, mit $3Å < RMSD \leq 4Å$ innerhalb und mit $4Å < RMSD < \infty Å$ oberhalb der sogenannten Zwielichtzone (Abb. 10). Innerhalb und oberhalb dieser Zone existierenden richtig-positiven Alignments sind nur schwer aus der Masse von falsch-positiven Alignments mit ähnlichen RMSD-Werten hervorzuheben. Die RMSD-Werte sollen nun so umgeformt werden, dass sie sinngemäß in etwa dieser Zoneneinteilung entsprechen. Eine mögliche Variante die RMSD zu normalisieren ist

$$NRMSD(K^A) = \frac{1}{1 + RMSD(K^A)} \tag{34}$$

(Abb. 10, schwarze Kurve). Diese Variante ist jedoch nicht zutreffend, da z.B. zwei sehr gut übereinstimmenden Strukturen mit einer $RMSD = 1.0Å$ lediglich mit $0.5 \hat{=} 50\%$ bewertet werden. Darüber hinaus konvergiert die Kurve für große RMSD-Werte, die ab $RMSD > 4Å$ unbedeutend werden, zu langsam gegen 0. Im Rahmen dieser Arbeit wird vorgeschlagen, die RMSD bzw. die WRMSD mittels einer sigmoiden logistischen Funktion

$$NRMSD(K^A) = \frac{1 + e^{-\lambda}}{1 + e^{RMSD(K^A) - \lambda}} \tag{35}$$

bzw.

$$NWRMSD(K^A) = \frac{1 + e^{-\lambda}}{1 + e^{WRMSD(K^A) - \lambda}} \tag{36}$$

zu normalisieren (Abb. 10, rot, grün, blau). Der Parameter λ triggert den Wendepunkt der Kurve. Je größer das λ, desto weiter verschiebt er sich nach rechts - desto höhere RMSD-

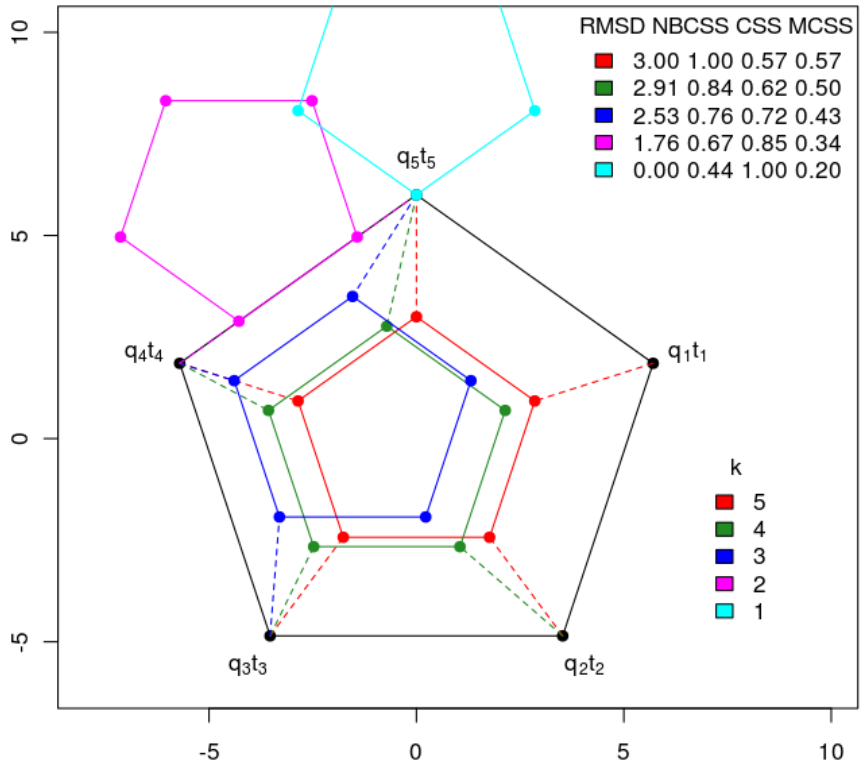


Abbildung 11: Überlagerung formgleicher Strukturen.

Werte werden besser bewertet. Die Wahl $\lambda = e = 2.718282$ ist empirisch und kann vom Benutzer frei gewählt werden. Die Normalized Root Mean Square Deviation (NRMSD)- bzw. NWRMSD-Werte können in dieser Form für die Berechnung von weiteren Deskriptoren verwendet werden. Die darauf basierenden Deskriptoren und ihre Bedeutung werden nun

anhand eines simplen Beispiels demonstriert. Abb. 11 zeigt eine Überlagerung zweier formgleichen Strukturen. Das kleinere regelmäßige Fünfeck, das eine schrittweise Bewegung erfährt, repräsentiert die QS. Die Bewegung der QS ist schrittweise farblich unterlegt. Das große schwarze regelmäßige Fünfeck, das in der Mitte des Bildes fixiert bleibt, ist die TS. Die Kanten der TS sind genau zweimal so groß wie die Kanten der QS. Die Eckpunkte der beiden Polygone können im übertragenen Sinne als Cα-Atome verstanden werden. Die Überlagerung der beiden Strukturen erfolgt absolut spezifisch, sodass q_1 nur t_1, q_2 nur t_2, ..., q_5 nur t_5 entsprechen darf. Unter dieser Voraussetzung existiert genau eine Alignmentkombination $K_5^A = ((q_1, t_1), \ldots, (q_5, t_5))$ der fünften Ordnung $k = |K_5^A| = 5$. Die Überlagerung der beiden Strukturen nach dieser Zuordnungsvorschrift ist in rot dargestellt. Die roten gestrichelten Linien symbolisieren die Distanzen, die zur Berechnung von $RMSD$ verwendet werden. Alle fünf Distanzen sind gleich und die $RMSD = 3.0\text{Å}$. Entfernt man einen Tupel (q_i, t_j) mit der größten Distanz d_{max}, dann kommt man zu einer möglichen Eltern-Alignmentkombination der Ordnung $k-1$. Das Entfernen des Tupels mit der größten Distanz führt zwangsläufig zu einer Verbesserung von RMSD und somit zu einer ähnlicheren Substruktur. Im dargestellten Beispiel wird das Tupel (q_1, t_1) aus K_5^A entfernt, sodass eine neue Alignmentkombination $K_4^A = ((q_2, t_2), \ldots, (q_5, t_5))$ der vierten Ordnung $k = 4$ mit $RMSD = 2.91\text{Å}$ entsteht. Diese Überlagerung ist in grün dargestellt. Entfernt man nach dem selben Prinzip die weiteren Tupel, bis die Alignmentkombination der kleinstmöglichen Ordnung $k = 1$ erreicht ist, dann hat man einen möglichen Pfad bewandert, dessen Alignmentkombinationen in einer Eltern-Kind-Beziehung stehen. Jede so ermittelte Alignmentkombination ist eine mögliche CS der beiden Strukturen. Abb. 12 stellt eine Überlagerung zweier formähnlichen Strukturen dar.

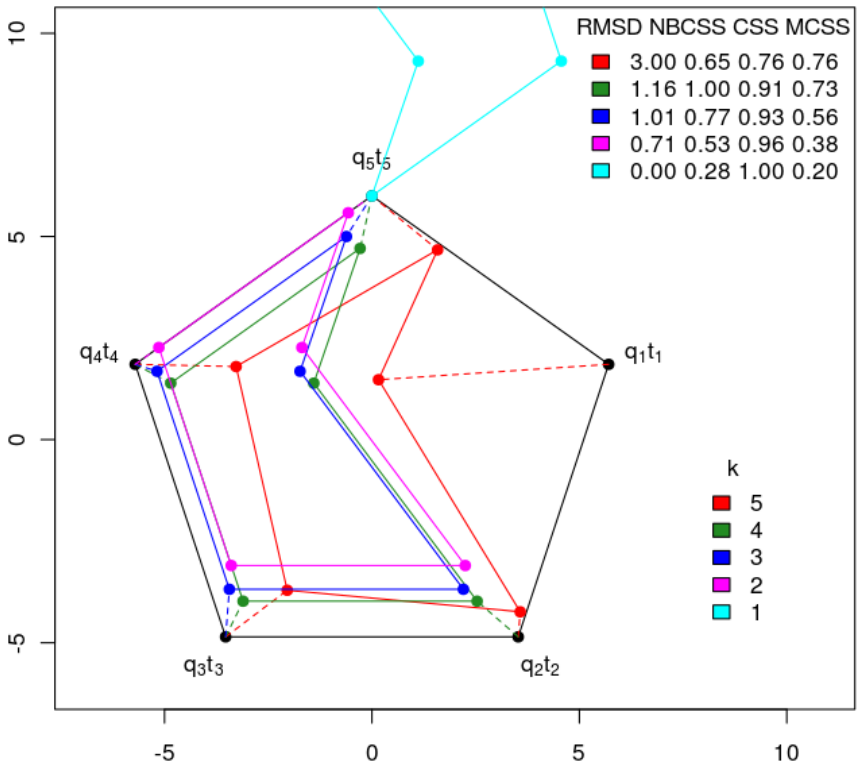


Abbildung 12: Überlagerung formähnlicher Strukturen.

Die Koordinaten der Eckpunkte der QS aus Abb. 11 sind so manipuliert, das sich dadurch zwar eine neue Distanzmatrix ergibt, aber die RMSD nach der Überlagerung mit der selben TS gleich bleibt ($RMSD = 3.0\text{Å}$). Die Vorgehensweise bei der Pfadwanderung bleibt die gleiche und basiert auf dem *top-down*-Prinzip. D.h. der Übergang von einer gemeinsamen

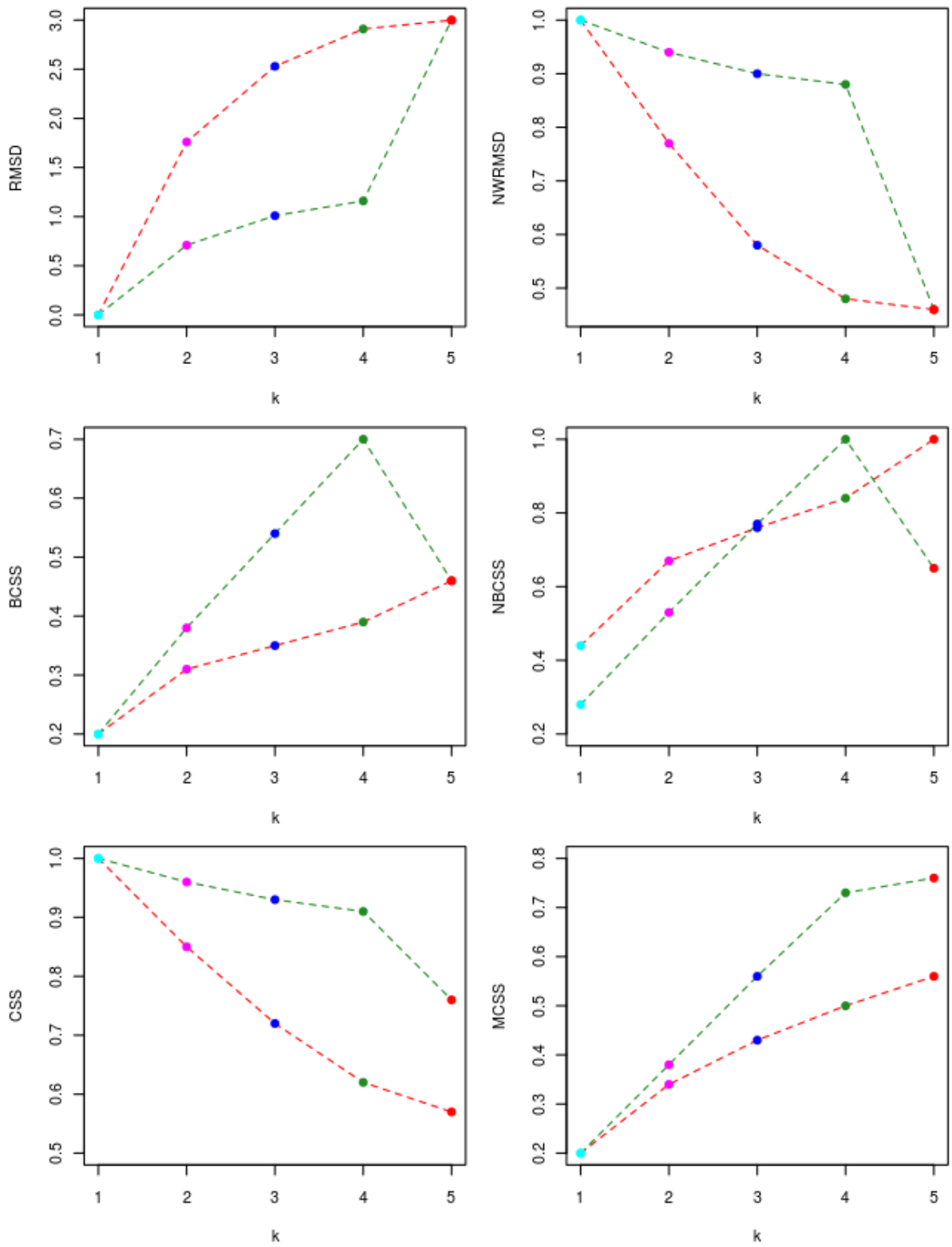


Abbildung 13: Deskriptoren & Pfade (*top-down*). Farbpunkte markieren die Deskriptorwerte der jeweiligen Substruktur der jeweiligen Ordnung k. Der jeweilige Pfad, bzw. die Übergänge von den Eltern- zu den Kind-Substrukturen (Ordnung $k = 5$ bis $k = 1$, *top-down*), können entlang der gestrichelten Linien (rot für Abb. 11, grün für Abb. 12) verfolgt werden.

Substruktur zu der nächstkleineren gemeinsamen Substruktur erfolgt unter Ausschluss ei-

nes einzigen Residuentupels mit der größten Distanz zwischen den korrespondierenden Eckpunkten. Ein solches Entfernen eines Residuentupels führt jedoch nicht zwangsläufig zu einer Substruktur, die für die Ordnung k das beste Alternativalignment ist. Dies ist der Grund, weshalb keine optimale Lösung garantiert werden kann. Für eine optimale Lösung müsste die Wahl der nächstkleineren gemeinsamen Substruktur auf der Auswertung aller möglichen Substrukturpfade basieren, was jedoch den Gegenstand der kombinatorischen Explosion ausmacht. Diese Suboptimalität der Heuristik wird dadurch minimiert, dass die Pfadwanderung nach dem *bottom-up*-Prinzip (Abs. 2.2.6.3) aufgebaut wird, wobei der Übergang zu den nächstgrößeren Substrukturen auf der Basis der besten Alternativalignments pro Ordnung k geschieht. Abb. 13 fasst die ermittelten Informationen über die überlagerten Strukturen zusammen. Die roten bzw. grünen gestrichelten Linien bilden das Geschehen in der Abb. 11 bzw. Abb. 12 ab. Die gefüllten Kreise markieren den jeweiligen Wert entsprechender Ordnung. Auf der Basis von RMSD und der pfadorientierten Betrachtung der gemeinsamen Substrukturen lassen sich die folgenden Deskriptoren ableiten:

RMSD In beiden Beispielen sind alle Query- und Target-Objekte einander zugeordnet. In beiden Fällen ist die MCS eine Alignmentkombination der fünften Ordnung (rot) mit der $RMSD = 3.0\mathring{A}$. Theoretisch existieren unendlich viele Möglichkeiten die Koordinaten der QS aus Abb. 11 so zu verändern, dass ihre RMSD nach der Überlagerung mit der TS gleich $3.0\mathring{A}$ bleibt. Diese Tatsache macht den Nachteil der alleinigen Verwendung von RMSD als Deskriptor deutlich. Die RMSD gibt aufgrund ihrer Mittelwertnatur keine Auskunft über die Ähnlichkeit der Substrukturen der beiden überlagerten Strukturen. Die Unterschiede in den RMSD-Werten ab der vierten Ordnung abwärts zeigen jedoch, dass die Substrukturen der beiden MCSs sich deutlich voneinander unterscheiden. Die Substrukturen der Ordnungen 2 bis 4 der MCS aus Abb. 12 sind den Substrukturen des Targets ähnlicher als die Substrukturen derselben Ordnungen der MCS aus Abb. 11.

NWRMSD Normalisierung der RMSD vereinfacht ihre Verwendung in den weiteren Berechnungen. In den oben genannten Beispielen aus Abb. 11 und Abb. 12 sind alle Tupelgewichte $RMSS(q_i, t_j) = 1$, sodass die Gewichtung aus der Gl. 33 herausfällt und

$$NRMSD(K_k^A) = NWRMSD(K_k^A)$$

für alle Alignmentkombinationen gilt.

BCSS Beim Vergleich zweier Strukturen stellt sich allgemein die Frage, welche der ermittelten CSs nun die bessere ist? Ist es z.B. die MCS? Oder eine kleinere CS mit der besseren RMSD? Es ist also eine Frage der Balance zwischen den Größen der beiden Strukturen und ihrer geometrischen und physiko-chemischen Übereinstimmung. In diesem Zusammenhang bewertet der Best Common Substructure Score (BCSS)

$$BCSS(K_k^A) = \frac{k}{min(|Q|, |T|)} \cdot NWRMSD(K_k^A) \tag{37}$$

jede Alignmentkombination. Alignmentkombination mit dem größten BCSS stellt im Rahmen von EpitopeMatch den Kern der Ähnlichkeit dar. Im Fall der Abb. 11 besitzt das größte CS den höchsten BCSS, $BCSS_{max} = BCSS(K_5^A) = 0.46$. Im Fall der Abb. 12 besitzt die zweitgrößte CS den höchsten BCSS, $BCSS_{max} = BCSS(K_4^A) = 0.7$. Die relativ flache Steigung der BCSSs im ersten Fall zeugt von einer gleichmäßigen Verteilung von Substrukturen mit mäßigen Ähnlichkeiten. Die steile Steigung der BCSSs der Ordnungen 1 bis 4 im zweiten Fall zeugt von einer gleichmäßigen Verteilung von Substrukturen mit hohen Ähnlichkeiten, sodass genau sie den Kern der Ähnlichkeit darstellen.

Der deutliche Abstieg des BCSS der fünften Ordnung lässt diese Alignmentkombination aus dem Rahmen springen. Die BCSSs der MCSs sind in beiden Fällen gleich 0.46. Ob nun die größere MCS im ersten Fall oder die kleinere CS im zweiten Fall die bessere ist, entscheidet je nach Bedarf der Benutzer selbst. Die Aufgabe von EpitopeMatch ist, diese Information an den Benutzer zu tragen.

NBCSS Die BCS, d.h. der Ähnlichkeitskern, ist die CS mit dem höchsten BCSS. Um diese Eigenschaft für den bequemeren Einsatz beim multiplen Strukturalignment zu vereinheitlichen, wird der BCSS als Normalized Best Common Substructure Score (NBCSS) normalisiert

$$NBCSS(K_k^A) = \frac{BCSS(K_k^A)}{BCSS_{max}} \tag{38}$$

So können alle BCSs aus paarweisen Alignments direkt über $NBCSS = 1.0$ gefiltert werden. In der Regel können alle CSs mit $0.9 \leqq NBCSS \leqq 1.0$ als Alternativalignments des Ähnlichkeitskerns betrachtet werden.

CSS Jede CS kann in Form eines Pfades (Abb. 13 (*top-down*), vgl. Abb. 15b (*bottom-up*))

$$P(K_k^A) = \{K_1^A, \ldots, K_k^A\} \tag{39}$$

dargestellt werden. Jedes Element des Pfades ist eine Alignmentkombination bzw. eine Substruktur, die mit der nächsten Alignmentkombination in einer Eltern-Kind-Beziehung steht. Der Unterschied zwischen zwei aufeinander folgenden Alignmentkombinationen ist ein Residuentupel. Die Anzahl der existierenden Pfade ist gleich der Anzahl der existierenden Alignmentkombinationen, d.h., jede Alignmentkombination ist ein Pfad. Eine solche Zerlegung einer Alignmentkombination in ihre Substrukturen erlaubt folgende Definition des Common Substructure Score (CSS)

$$CSS(P(K_k^A)) = \frac{2}{|P(K_k^A)|^2 + |P(K_k^A)|} \cdot \sum_{k=1}^{|P(K_k^A)|} k \cdot NWRMSD(K_k^A) \tag{40}$$

Bewertung einer CS mit diesem Deskriptor impliziert somit die Bewertung jeder Substruktur des gesamten CS-Pfades. Während die MCSs der beiden Fälle aus Abb. 11 und Abb. 12 mit

$$NWRMSD(K_5^A) = 0.46 \hat{=} 46\%$$

gleichwertig sind, werden sie mit

$$CSS(P(K_5^A)) = 0.57 \hat{=} 57\%$$

im ersten Fall und mit

$$CSS(P(K_5^A)) = 0.76 \hat{=} 76\%$$

im zweiten Fall deutlich voneinander unterschieden. Trotz des Ausreißers wird die zweite Struktur als ähnlicher bewertet, da sie über eine größere Anzahl von ähnlicheren Substrukturen verfügt.

RGD Während der CSS mit NWRMSD auch die physiko-chemische Ähnlichkeit der Residuen SSIM in die Bewertung der Ähnlichkeit der gemeinsamen Substruktur CS injiziert, kann unter Ausschluss der SSIM die RiGiDity (RGD) gemessen werden

$$RGD(P(K_k^A)) = \frac{2}{|P(K_k^A)|^2 + |P(K_k^A)|} \cdot \sum_{k=1}^{|P(K_k^A)|} k \cdot NRMSD(K_k^A) \tag{41}$$

Die Messung der Starrheit bzw. der Rigidität RGD erfolgt also anhand der Auswertung der geometrischen Ähnlichkeit NRMSD aller Substrukturen einer CS, auf denen sie entlang des definierten Pfades $P(K_k^A)$ aufbaut. Wenn alle Tupelgewichte der beiden zu vergleichenden Strukturen $RMSS(q_i, t_j) = 1$ sind, d.h. die physikochemische Ähnlichkeit der einander zugeordneten Residuen identisch ist, dann ist $NRMSD(K_k^A) = NWRMSD(K_k^A)$ und somit $RGD(P(K_k^A)) = CSS(P(K_k^A))$. Somit beträgt die RGD der CS im ersten Fall

$$RGD(P(K_k^A)) = CSS(P(K_5^A)) = 0.57\hat{=}57\%$$

und die RGD der CS im zweiten Fall.

$$RGD(P(K_k^A)) = CSS(P(K_5^A)) = 0.76\hat{=}76\%$$

Anhand der alleinigen Auswertung der RMSD bzw. der NRMSD, die in den beiden Fällen $RMSD(K_5^A) = 3.0\mathring{A}$ bzw. $NRMSD(K_5^A) = 0.46\hat{=}46\%$ sind, würde man keine Unterschiede der CSs feststellen können. Die Messung der RGD zeigt jedoch, dass die CS im zweiten Fall über mehr starre Anteile verfügt als die CS im ersten Fall, die von den größeren Bewegungen geprägt ist. Abhängig von dem gewählten Template kann auf diese Weise die RGD der Cα-Atome, des BACKBONE(N,Cα,C′,O) und die ALLATOMS-RGD explizit gemessen werden.

FLX Die FLeXibility (FLX) kann auf diesem Hintergrund als

$$FLX(P(K_k^A)) = 1 - RGD(P(K_5^A)) \tag{42}$$

ausgedrückt werden, womit die Flexibilität der CS im ersten Fall größer

$$1 - 0.57\hat{=}43\% > 24\%\hat{=}1 - 0.76$$

als die Flexibilität der CS im zweiten Fall ist.

IDENT Zusätzlich informiert die identity (IDENT)

$$IDENT(K^A) = \frac{\sum_{k=1}^{|K^A|} \begin{cases} 1 & CODE(Q_k) = CODE(T_k) \\ 0 & CODE(Q_k) \neq CODE(T_k) \end{cases}}{|K^A|} \tag{43}$$

über die Anzahl der Residuentupel mit identischen Query- und Target-Residuen.

SSIM Die SSIM

$$SSIM(K^A) = \frac{\sum_{k=1}^{|K^A|} ss(RT_k^{QT})}{|K^A|} \tag{44}$$

gibt in Abhängigkeit von der gewählten Substitutionsmatrix die mittlere Substitutionsähnlichkeit einer CS an.

QMCSS Der Query Maximum Common Substructure Score (QMCSS) beschreibt

$$QMCSS(P(K_k^A)) = \frac{|P(K_k^A)|}{|Q|} \cdot CSS(P(K_k^A)) \tag{45}$$

das Verhältnis der Ähnlichkeit einer CS zu der Größe der QS.

TMCSS Analog beschreibt der Target Maximum Common Substructure Score (TMCSS)

$$TMCSS(P(K_k^A)) = \frac{|P(K_k^A)|}{|T|} \cdot CSS(P(K_k^A)) \tag{46}$$

das Verhältnis der Ähnlichkeit einer CS zu der Größe der TS.

QTMCSS Der Query-Target Maximum Common Substructure Score (QTMCSS) drückt

$$QTMCSS(P(K_k^A)) = \frac{2 \cdot |P(K_k^A)|}{|Q| + |T|} \cdot CSS(P(K_k^A)) \tag{47}$$

die Beziehung der Größen der beiden Strukturen zu ihrer Ähnlichkeit aus.

MCSS Die größte, im Rahmen der gesetzten Parameter ermittelte CS, ist die MCS. Der Maximum Common Substructure Score (MCSS)

$$MCSS(P(K_k^A)) = \frac{|P(K_k^A)|}{min(|Q|, |T|)} \cdot CSS(P(K_k^A)) \tag{48}$$

ist die direkte Beziehung zwischen der Ähnlichkeit einer Substruktur CSS und der maximal möglichen Größe einer CS $min(|Q|, |T|)$. Ferner gilt

$$MCSS(P(K_k^A)) = \begin{cases} QMCSS(P(K_k^A)) & |Q| \leqq |T| \\ TMCSS(P(K_k^A)) & |Q| > |T| \end{cases} \tag{49}$$

COMPL Die COMPL einer CS wird wie folgt berechnet

$$COMPL(P(K_k^A)) = \frac{|P(K_k^A)|}{min(|Q|, |T|)} = \frac{MCSS(P(K_k^A))}{CSS(P(K_k^A))} \tag{50}$$

Die oben aufgeführten Deskriptoren lassen sich bezüglich der Größe der gesuchten Substruktur in zwei Gruppen aufteilen: die *größenunabhängigen* Deskriptoren RMSD, NRMSD, NWRMSD, CSS, RGD, FLX, IDENT, SSIM; und die *größenabhängigen* Deskriptoren BCSS, NBCSS, QMCSS, TMCSS, QTMCSS, MCSS, COMPL.

DRMSD Die mittels RMSD gemessene Abweichung der Konformation pro Substruktur CS liefert angefangen mit der minimalen Definition des Rückgrats (Cα-Template) bis hin zu der maximalen Betrachtung der Residuen (AllAtoms-Template) einen Hinweis auf die Flexibilität der CS. Bleiben die Konformationsänderungen auf die Verschiebungen der Rotamere und kleinere Rückgratkonformationsänderungen beschränkt, d.h. unterhalb der Zwielichtzone ($\leqq 3.0\mathring{A}$), so kann die entsprechende Substruktur als ein Ganzes betrachtet und verwertet werden. Bewegt sich die RMSD einer CS innerhalb oder außerhalb der Zwielichtzone ($> 3.0\mathring{A}$), dann handelt es sich im Fall eines richtig-positiven Alignments oft um die mitgemessenen Domänenbewegungen, sodass die entsprechende CS in mehrere CSs gespalten werden kann, die sich wieder unterhalb der Zwielichtzone einfinden. Die Messung der Domänenbewegungen, die meist mit den relativ großen Verschiebungen der in der Regel starr bleibenden Domänen verbunden sind, wird auf diese Weise aus der Bewertung der geometrischen Ähnlichkeit herausgenommen und auf die Definition der Hinge-Bending-Regionen (Abs. 2.2.7)

verlagert, deren Residuen oft die ungewöhnlichen Torsionswinkel [137, 57] einnehmen. Die Domain-Scores der größenunabhängigen Deskriptoren können analog zu der Domain Root Mean Square Deviation (DRMSD)

$$DRMSD(CS_1, ..., CS_N) = \frac{\sum_{k=1}^{N} |CS_k| \cdot RMSD(CS_k)}{\sum_{k=1}^{N} |CS_k|} \tag{51}$$

berechnet werden.

DMCSS Die Domain-Scores der größenabhängigen Deskriptoren können analog zu dem Domain Maximum Common Substructure Score (DMCSS)

$$DMCSS(CS_1, ..., CS_N) = \sum_{k=1}^{N} MCSS(CS_k) \tag{52}$$

berechnet werden.

Die Gesamtheit der oben definierten Deskriptoren deckt das Spektrum der Ähnlichkeitsmerkmale der gemeinsamen Substrukturen für die Suche nach dem optimalen Strukturalignment hinreichend ab. Betrachtung einzelner Deskriptoren erlaubt die Konzentration auf die jeweilige Merkmalsausprägung des Ähnlichkeitsspektrums. In den folgenden Kapiteln wird demonstriert, wie die Deskriptoren bei den qualitativen (2.3.1.3) und quantitativen (2.3.1.4) Messungen der Ähnlichkeit ihren Einsatz finden.

2.2.6.2 *Gemeinsame Substruktur*

Das Beispiel aus Abb. 8a wird hier zur Anschauung weiter geführt. Jede Alignmentkombination $K^A \in A$ (Abb. 14, grün) mit $|Q| = 141$ (Struktur $2DN2$ Kette A) und $|T| = 146$

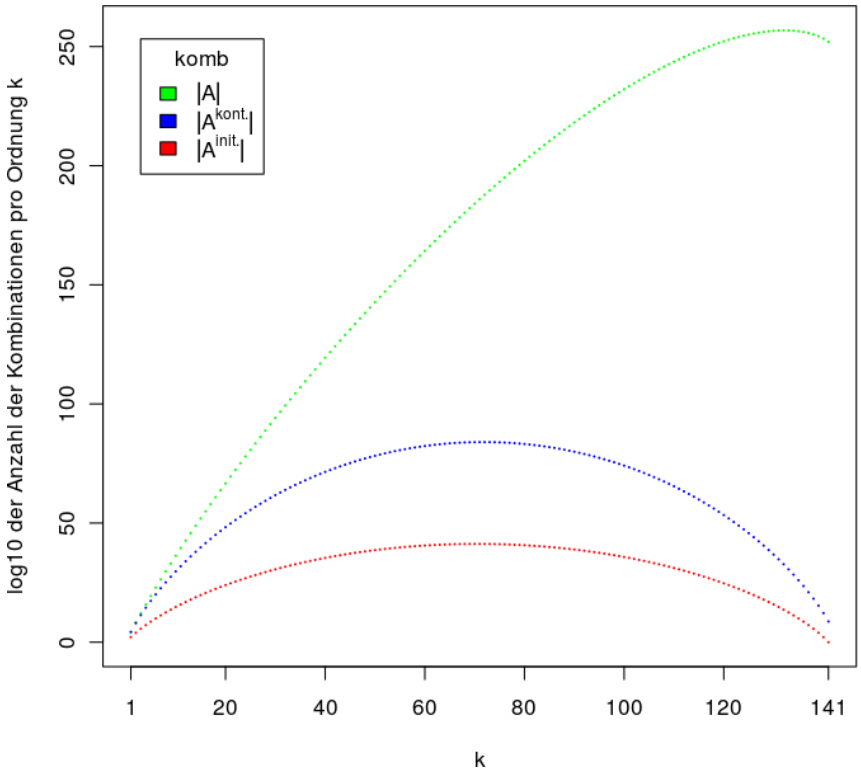


Abbildung 14: Reduktion der kombinatorischen Explosion.

(Struktur $2DN2$ Kette B) ist eine potentielle CS der beiden Strukturen. Im diskontinuierlichen Fall gibt es also insgesamt

$$|A| = \sum_{k=1}^{141} \binom{141}{k} \binom{146}{k} k! = 3.678287 \cdot 10^{257}$$

CSs der Ordnungen $1 \leq k \leq 141$, davon sind allein $9.791643 \cdot 10^{251}$ Kombinationen der Ordnung $k = 141$. Im kontinuierlichen Fall (Abb. 14, blau) existieren

$$|A^{kont.}| = \sum_{k=1}^{141} \binom{141}{k} \binom{146}{k} = 1.120415 \cdot 10^{85}$$

CSs der Ordnungen $1 \leq k \leq 141$, davon sind $5.158536 \cdot 10^{8}$ Kombinationen der Ordnung $k = 141$. Zum Vergleich, das Universum enthält schätzungsweise $1.57 \cdot 10^{79}$ Protonen. Berechnung der Deskriptoren für jede mögliche CS ist also höchst ineffizient.

Das Ergebnis der ersten Algorithmusstufe ist die initiale Alignmentkombination. Im diskontinuierlichen Fall ist sie eine $\binom{141}{141} = 1$ von $\binom{146}{141} \cdot 141! = 9.791643 \cdot 10^{251}$ Kombinationen der Ordnung $k = 141$ (Abb. 14, rot). Im Idealfall, z.B. beim Vergleichen von identischen oder nahezu identischen Strukturen, wären alle 141 Residuentupel RT^{QT} der initialen Alignmentkombination richtig-positiv

$$|CS(RT^{QT})| = |CS(RT^{QT,TP})|$$

Diese CS wäre gleichermaßen die MCS und die BCS der beiden Strukturen und würde als Antwort auf die Frage nach Ähnlichkeit zwischen den beiden Strukturen absolut genügen. Im Regelfall setzt sich die Menge der Residuentupel der initialen Alignmentkombination aus den richtig-positiven $RT^{QT,TP}$ und den falsch-positiven $RT^{QT,FP}$ Residuentupel zusammen. Je unähnlicher sich die beiden Strukturen sind, desto weniger $RT^{QT,TP}$ und desto mehr $RT^{QT,FP}$ sind in dieser Menge enthalten. Aus den Tupeln der initialen Alignmentkombination lassen sich insgesamt

$$|A^{init.}| = \sum_{k=1}^{141} \binom{141}{k} = 2.787593 \cdot 10^{42}$$

CSs der Ordnungen $1 \leq k \leq min(|Q|, |T|)$ kombinieren (Abb. 14, rot). Die Anzahl der zu berechnenden CSs ist für den Vergleich der Proteine mit der durchschnittlichen Größe immer noch zu groß und muss weiterhin reduziert werden.

2.2.6.3 *Bottom-Up*

Eine Superposition der beiden Strukturen anhand der Zuordnungsvorschrift aus der initialen Alignmentkombination ist kein aussagekräftiger Ausgangspunkt, von dem aus die falsch-positiven Residuentupel aussortiert und die richtig-positiven CSs konstruiert werden können. Die Gründe:

- Ist die Anzahl der richtig-positiven Residuentupel einer CS kleiner als 50 Prozent $|CS(RT^{QT,TP})| < 50\%$, dann ist die Verzerrung der Überlagerung der richtig-positiven Residuen so groß, dass sie von den falsch-positiven Residuen nicht effektiv unterschieden werden können.

- Bewegt man sich angefangen mit der initialen Alignmentkombination der Ordnung

$$k = min(|Q|, |T|)$$

in Richtung $k = 1$ (Abb. 14, rot), so hat man als erstes die Menge der größten CSs auszuwerten. Die Verteilung der Werte (Abb. 14, rot) um die mittlere Ordnung $k = 71$ ist symmetrisch. Dementsprechend ist z.B. die Anzahl der Kombinationen $\binom{141}{2} = \binom{141}{139} = 9870$ der Ordnungen $k = 2$ und $k = 139$ gleich. Es liegt auf der Hand, dass die Berechnung von 9870 Überlagerungen bzw. Deskriptoren der Kombinationen mit jeweils 2 Residuentupeln ca. $\frac{139}{2} = 69.5$ Mal schneller ist, als die Berechnung von 9870 Überlagerungen der Kombinationen mit jeweils 139 Residuentupeln.

Diese Tatsachen zeugen von der Ineffizienz einer *top-down*-Durchwanderung des Kombinationsraums.

Demgegenüber steht die *bottom-up*-Methode. Die erste Algorithmusstufe liefert in Form von der initialen Alignmentkombination zunächst die einzige und zugleich die größte bekannte CS der Ordnung $k = min(|Q|, |T|)$, mit einer Mischung aus richtig- und falsch-positiven Residuentupeln. Zugleich, ist jedes Residuentupel der initialen Alignmentkombination die kleinste bekannte, entweder richtig- oder falsch-positive CS der Ordnung $k = 1$. Nun stelle man sich die Ordnung $k = 1$ aus Abb. 14 als ein Stapel vor (Abb. 15a), auf dem alle $\binom{141}{1} = 141$ CSs der ersten Ordnung liegen. Jede CS ist sowohl eine Alignmentkombination K_1^A als auch ein Pfad (Abb. 15b) der Alignmentkombinationen $P(K_1^A) = \{K_1^A\}$. Alle $\binom{141}{1} = 141$ CSs werden nun superpositioniert. Nach der jeweiligen Superposition erfolgt die Berechnung der Deskriptoren. Die CSs werden nach dem Deskriptor $CSS(P(K_1^A))$ auf dem jeweiligen Stapel absteigend sortiert. Somit ist die geometrische und die physiko-chemische Ähnlichkeit der einzelnen Residuentupel, bzw. der kleinsten Substrukturen, aus der initialen Alignmentkombination $K_{141}^{A,init.}$ bekannt. In welchem Zusammenhang die Ähnlichkeiten der einzelnen Residuen der beiden Strukturen stehen, beantwortet das kombinatorische Resampling. Dieses gliedert sich in zwei Abschnitte: Suche nach der BCS; und Suche nach der MCS.

2.2.6.4 *Beste gemeinsame Substruktur*

Die Suche nach der BCS nutzt als erstes die Residuentupel der initialen Alignmentkombination. Für die zweite Ordnung $k = 2$ lassen sich $\binom{141}{2} = 9870$ CSs generieren. Jede neue Alignmentkombination der zweiten Ordnung K_2^A steht mit der entsprechenden Alignmentkombination der ersten Ordnung K_1^A in einer Eltern-Kind-Beziehung, sodass jede neue CS ein Pfad (Abb. 15b) $P(K_2^A) = \{K_1^A, K_2^A\}$ der Länge $|P(K_2^A)| = 2$ ist. Auch die Alignmentkombinationen der zweiten Ordnung kann man sich auf dem Stapel (Abb. 15a) liegend und nach dem Deskriptor $CSS(P(K_2^A))$ absteigend sortiert vorstellen. Interessant für die weitere Betrachtung sind in diesem Fall nur die CSs mit den höchsten $CSS(P(K_2^A))$. Zwar werden alle CSs der zweiten Ordnung erzeugt, es ist jedoch weder effizient noch notwendig alle CSs zu speichern. Wie viele CSs pro Ordnung k während der Suche nach BCS gespeichert werden, gibt die

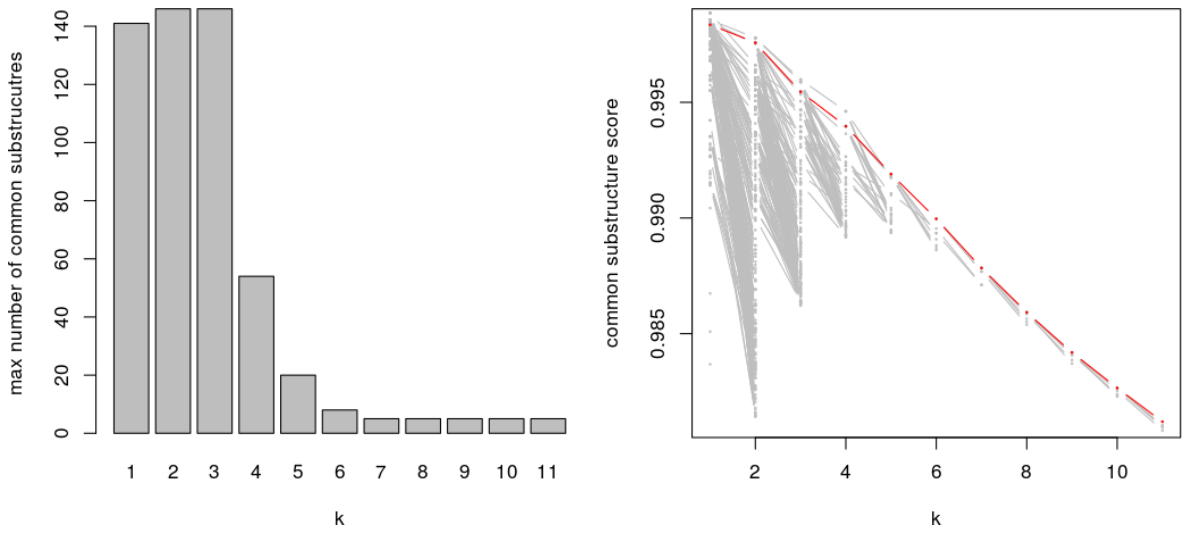
BCST Best Common Substructure Threshold (BCST) an

$$BCST_k = \begin{cases} |Q| & k = 1 \\ |T| & |T| \cdot e^{trigger-k} > |T| \\ BCST_{min} & |T| \cdot e^{trigger-k} < BCST_{min} \\ |T| \cdot e^{trigger-k} & sonst \end{cases} \tag{53}$$

mit

$1 \leq k \leq \lvert Q \rvert$	Mögliche Ordnungen.
$1 \leq trigger \leq \lvert Q \rvert$	Verteilungstrigger für die ersten Ordnungen.
$BCST_{min}$	Threshold für die minimale Anzahl von CSs.

Abb. 15a zeigt Die Anzahl der CSs pro Ordnung k für $\lvert Q \rvert = 141$, $\lvert T \rvert = 146$, $trigger = 3$



(a) Maximale Anzahl der gespeicherten CSs pro Ordnung k. Ein Stapel pro Ordnung k. Die Ordnungen 12 bis 141 sind in beiden Abbildungen nicht enthalten. Die CSs sind nach dem CSS (rechts) auf dem jeweiligen Stapel (Ordnung k) absteigend sortiert.

(b) Von jeder gespeicherten CS der Ordnung k wird eine CS der Ordnung $k+1$ abgeleitet. Die Pfade (Linien) resultieren in einer oder mehreren MCSs. Die beste MCS (rot) baut nicht unbedingt auf den besten CSs (Ordnungen 1 bis 4, grau) auf.

Abbildung 15: Stapel & Pfade (*bottom-up*).

und dem benutzerdefinierten Parameter $BCST_{min} = 5$. Die erste Ordnung enthält maximal $\lvert Q \rvert$ CSs, die aus der initialen Alignmentkombination resultieren. Je kleiner die Substrukturen, desto häufiger treten die Ähnlichkeiten zwischen ihnen auf bzw. desto zahlreicher ist das Vorkommen einander sehr ähnlicher Substrukturen. Die Ordnungen zwei und drei sind so getriggert, dass pro Ordnung maximal $\lvert T \rvert$ CSs gespeichert werden - eine ausreichende Anzahl von Kombinationen, um auf ihrer Grundlage die Kombinationen der nächsten Ordnung zu erzeugen. Mit der wachsenden Größe der Substrukturen wachsen auch die Unterschiede in ihrer Ähnlichkeit. Die unähnlichen Pfade divergieren von den besten Pfaden an der Spitze des Stapels mit der wachsenden Größe deutlich genug und können anhand der Unterschiede ihrer CSSs früh genug aussortiert werden. Der Kern der Ähnlichkeit ist in der Regel ab der Ordnung fünf oder sechs identifiziert. Ab der Ordnung sieben werden maximal $BCST_{min} = 5$ CSs gespeichert, die in der Regel Permutationen der richtig-positiven Residuentupel der jeweiligen Ordnung sind. Im Idealfall (vergleichen von identischen oder nahezu identischen Strukturen), würde die Speicherung von maximal einer CS pro Ordnung $BCST = \{1, \ldots, 1\}$ völlig genügen. Damit wäre der gesamte Kombinationsraum unter der roten Kurve (Abb. 14) auf lediglich

$$BCST_{141} + \sum_{k=1}^{140} BCST_k \cdot (141 - k) = 1 + \sum_{k=1}^{140} 1 \cdot (141 - k) = 9871$$

zu berechnenden Kombinationen reduziert. Im Regelfall ist nicht jede Kombination unter der roten Kurve richtig-positiv. Aus diesem Grund sollte eine gewisse Kombina-

tionsfreiheit pro Ordnung gegeben sein, um das Hineinmanövrieren in die Sackgassen, durch das Erweitern der Pfade um die falsch-positiven Residuentupel, zu vermeiden. Für das Beispiel aus Abb. 15 mit $BCST = \{141, 146, 146, 54, 20, 8, 5, \ldots, 5\}$ werden höchstens

$$BCST_{|Q|} + \sum_{k=1}^{|Q|-1} BCST_k \cdot (|Q| - k) = 1.1661 \cdot 10^5 \lll 2.787593 \cdot 10^{42}$$

Alignmentkombinationen generiert. Davon werden höchstens

$$\sum_{k=1}^{141} BCST_k = 1190$$

als CSs gespeichert. Die Speicherung der CSs auf dem Stapel der jeweiligen Ordnung k erfolgt absteigend sortiert nach dem CSS. Die Anzahl n der CSs pro Ordnung k ist

$$1 \leqq n \leqq BCST_k$$

sodass jede CS der Ordnung k mit $CS_{k,n}$ adressiert werden kann. Die beste CS der Ordnung k ist die $CS_{k,1}$ mit dem größten CSS. Die schlechteste CS der Ordnung k ist die $CS_{k,BCST_k}$ mit dem kleinsten CSS. Eine CS wird gespeichert wenn die maximale Anzahl $BCST_k$ der erlaubten Speicherungen pro Ordnung k noch nicht erreicht ist

$$n < BCST_k$$

oder wenn sie erreicht ist

$$n = BCST_k \wedge CSS(CS_{k,n}) > CSS(CS_{k,BCST_k})$$

und der CSS der neuen CS besser als der CSS der schlechtesten gespeicherten CS ist. Im zweiten Fall wird die zuvor gespeicherte $CS_{k,BCST_k}$ vom Stapel entfernt. Die beiden Deskriptoren CSS und BCSS implizieren die geometrische und physiko-chemische Ähnlichkeit einer CS - beide Größen berechnen sich auf der Grundlage von NWRMSD. Der CSS ist unabhängig von der Größe der jeweiligen CS und sorgt für den lokalen Vergleich der CSs pro Ordnung k.

BCSST Der BCSS hingegen bewertet mit der NWRMSD der jeweiligen CS das Verhältnis der Ordnung k zu der maximal möglichen Größe einer CS $min(|Q|, |T|)$. Er wird als globaler Vergleichsdeskriptor für alle Ordnungen k verwendet und dient der Findung der BCS. Der benutzerdefinierte Parameter Best Common Substructure Score Threshold (BCSST) gibt an, wie stark sich die neu gebildeten CSs von der aktuellen BCS unterscheiden dürfen. Eine BCS ist eine CS mit dem größten BCSS und bildet im Rahmen von EpitopeMatch den Ähnlichkeitskern der beiden Strukturen ab. Eine CS wird gespeichert wenn

$$\frac{BCSS(CS)}{BCSS(BCS)} \geqq BCSST$$

das Verhältnis ihres BCSS zu dem BCSS der besten gespeicherten gemeinsamen Substruktur größer oder gleich BCSST ist. Dieser Parameter schließt also alle gemeinsame Substrukturen aus, die weniger als 50% Ähnlichkeit mit dem Ähnlichkeitskern der beiden Strukturen besitzen. Die Standardeinstellung des BCSST ist $BCSST = 0.5$.

RMSDT Die CSs können zusätzlich über den Parameter Root Mean Square Deviation Threshold (RMSDT) aussortiert werden. Eine CS wird gespeichert wenn

$$RMSD(CS) \leqq RMSDT$$

ist. Die Standardeinstellung des RMSDT ist $RMSDT = 3.0\mathring{A}$. Die größeren Werte würden z.B. das Vergleichen von Strukturen mit größeren konformationellen Abweichungen bzw. inklusive Domänenverschiebungen zulassen.

Hat die BCS die maximal mögliche Größe erreicht

$$|BCS| = min(|Q|, |T|)$$

so bricht der Algorithmus ab. In diesem Fall ist die BCS = MCS. Abb. 16 demonstriert die

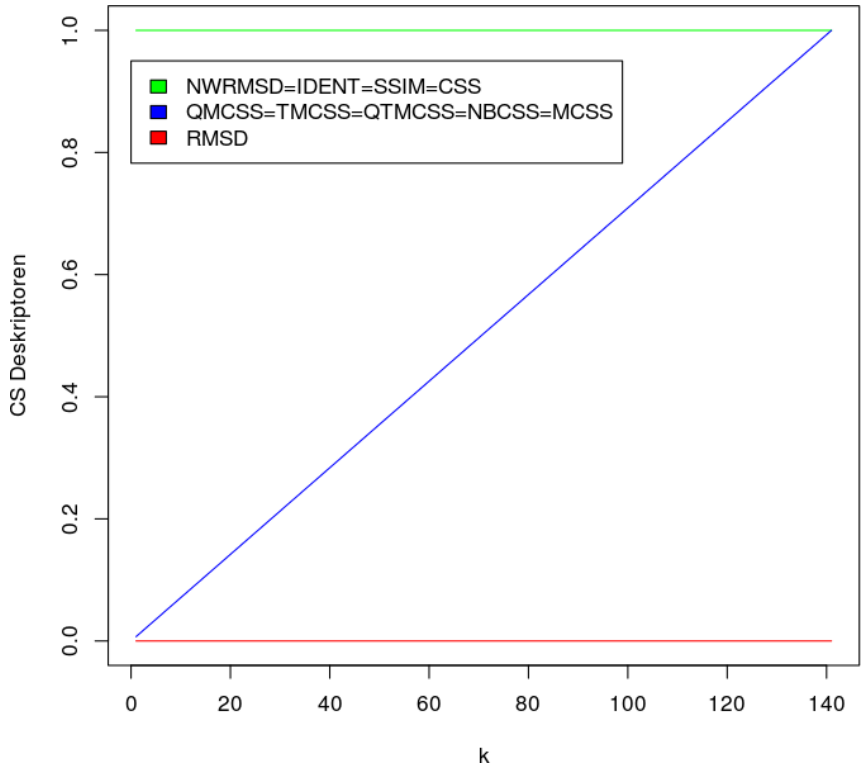


Abbildung 16: Verteilung der CS-Deskriptoren im Idealfall.

Verteilung der Deskriptorwerte der jeweiligen CS der Ordnung k nach dem Vergleichen der Kette A aus $2DN2$ mit sich selbst. Die Werte der von den CS-Größen unabhängigen Deskriptoren RMSD, NWRMSD, IDENT, SSIM und CSS sind gleich bei allen ermittelten CSs. Die Werte der von den CS-Größen abhängigen Deskriptoren QMCSS, TMCSS, QTMCSS, NBCSS und MCSS steigen proportional zu der Größe der CSs. Lediglich die größte CS, der Ordnung $k = 141$, ist im Idealfall von entscheidenden Interesse. Die BCS erreicht die maximal mögliche Größe nur beim Vergleichen von sehr ähnlichen Strukturen. In der Regel, d.h. beim Vergleichen homologer Strukturen, ist der Ähnlichkeitskern eine CS der Ordnung $k < min(|Q|, |T|)$

$$|BCS| < min(|Q|, |T|)$$

In diesem Fall sorgen die Parameter BCSST und RMSDT für die Terminiertheit des kombinatorischen Resampling. Die Suche nach der BCS bricht ab, wenn keine neue, nächsthöhere

Ordnung k gebildet werden kann. Somit ist das kombinatorische Potenzial der Residuentupel der initialen Alignmentkombination erschöpft. Die initiale Alignmentkombination aus dem Vergleich der Ketten A (Query) und B (Target) (beide aus $2DN2$) liefert im Fall *invariant*, gewichtet mit sigmoid normalisierten BLOSUM62-Matrix (Abb. 6, blau), 131 richtig- und 9 falsch-positive Residuentupel.

Abb. 17 zeigt die Deskriptorwerte der insgesamt 1145 gespeicherten CSs, die sich im Rahmen der gewählten Threshold-Parameter $BCST_{min} = 5$, $BCSST = 0.5$ und $RMSDT = 3.0\mathring{A}$ ergeben haben. Dem vorläufig gefundenen Ähnlichkeitskern BCS entspricht eine der fünf

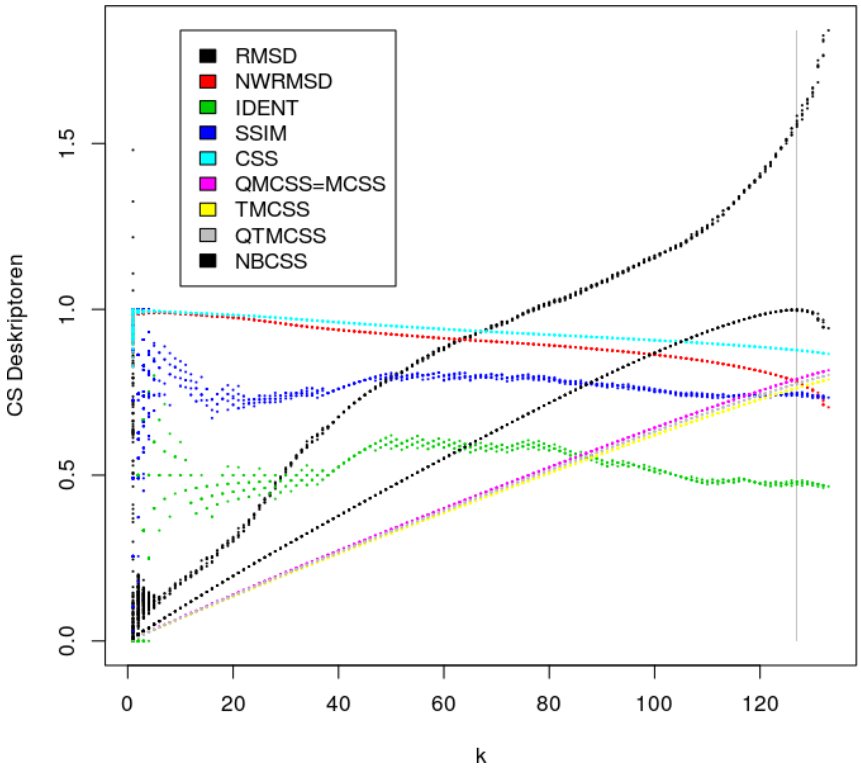


Abbildung 17: Deskriptoren der initialen CSs. Die graue Linie symbolisiert die BCS.

Alignmentkombinationen K^A_{127} der Ordnung $k = 127$ mit $NBCSS(CS) = 1.0$. Die größte ermittelte Alignmentkombination K^A_{133} der Ordnung $k = 133$ ist die vorläufige MCS. Tab. 2 fasst die Deskriptorwerte der gespeicherten Alignmentkombinationen zusammen.

COMPS	ATOMS	RMSD	NWRMSD	IDENT	SSIM	CSS	QMCSS	TMCSS	QTMCSS	NBCSS	MCSS
127	830	1.551	0.782	0.472	0.738	0.877	0.79	0.763	0.776	1.0	0.79
133	874	1.841	0.704	0.466	0.73	0.865	0.816	0.788	0.802	0.943	0.816

Tabelle 2: Deskriptoren der initialen BCS und MCS.

Die senkrechte Linie (Abb. 17) markiert die Ordnung der BCS. Unter den gespeicherten CSs existieren 39 CSs, die sich nur im Tausendstelbereich von der BCS unterscheiden $1.0 - NBCSS(CS) < 0.01$. Diese Alignmentkombinationen sind über die Ordnungen $122 \leqq k \leqq 130$ Verteilt und können als alternativen BCSs gesehen werden. Betrachtet man speziell die Eltern- und Kind-Kombinationen der BCS, so stellt man fest, dass die BCS der Punkt ist, an dem die entgegengesetzt gerichteten Eigenschaften der Eltern (kleinere Substruktur; bessere RMSD) und der Kinder (größere Substruktur; schlechtere RMSD) ausbalanciert sind. Abb. 18 demonstriert die BCS, die allein anhand von Residuentupel der initialen Alignmentkombination ermittelt wurde. Die Superposition erfolgte anhand von 830

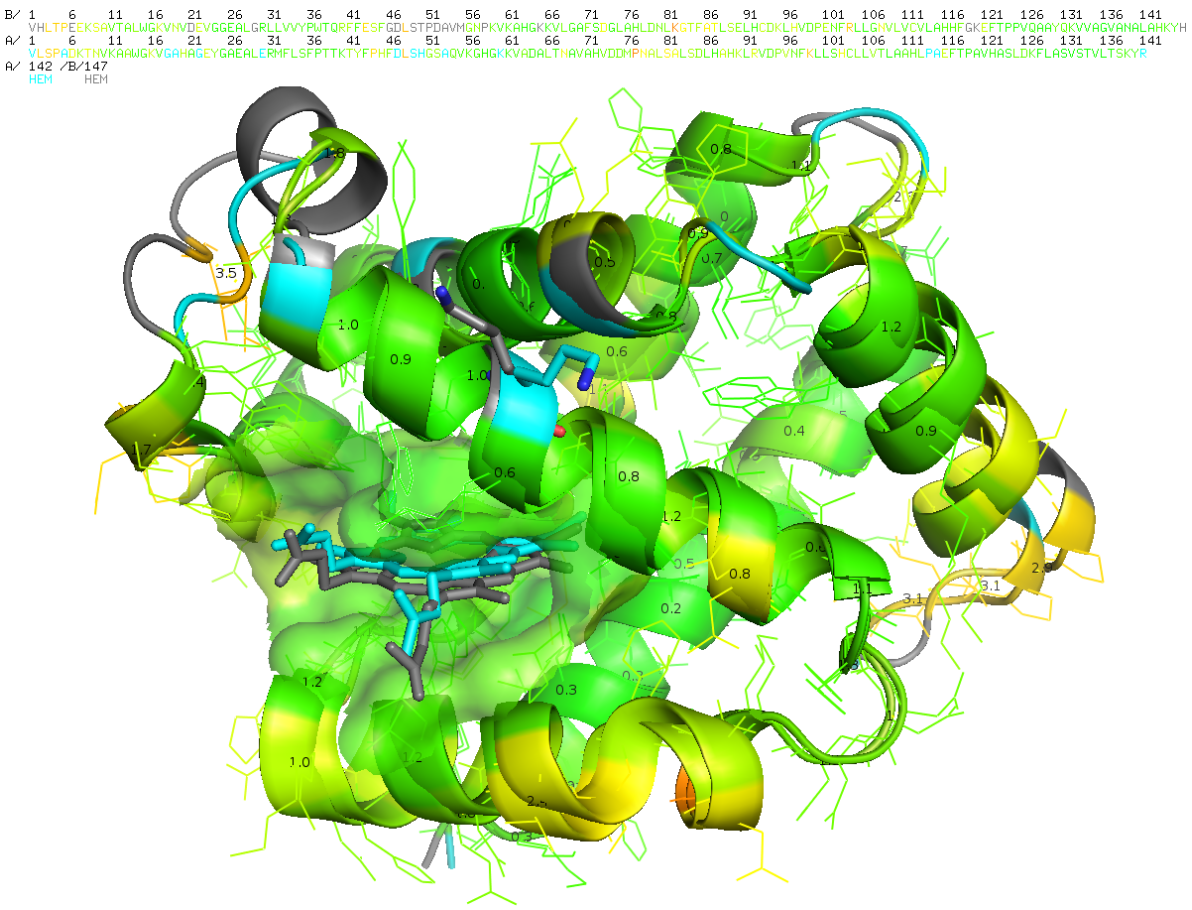


Abbildung 18: Die initale BCS. Die Färbung der korrespondierenden Residuen erfolgt nach der $NWRMSD(RT_{qt}^{QT})$ (min max) des jeweiligen Residuentupels.

Atomen (AllAtoms, Abb. 5) der 127 einander zugeordneten Residuen (Tab. 2). Einander nicht zugeordneten Residuen sind cyan (QS) und grau (TS). Die Färbung der zugeordneten Residuen wird weiter unten erläutert. Die Zahlen in schwarz geben die Entfernungen in $\mathring{A}$ zwischen den Cα-Atomen der zugeordneten Residuen an. Es ist deutlich zu sehen, dass die Regionen mit den Konformationsänderungen (Loops, Turns, Rotamere, ...), die aus dem Gesamtbild der konformationellen Übereinstimmungen herausfallen, unzugeordnet bleiben.

Während alle 127 Residuentupel der BCS richtig-positiv sind, besteht die MCS aus allen 131 richtig-positiven Residuentupel und den 2 von 9 falsch-positiven Residuentupel der initialen Alignmentkombination (Tab. 2, components (COMPS) = 133). Die beiden Strukturen verfügen jedoch über 139 richtig-positiven Zuordnungen. Ihre Ermittlung übernimmt das BCS-Resampling:

1. Transformiere alle Residuentupel RT_{qt}^{QT} aus der Scoringmatrix SM^{QT} anhand der Transformationsdaten der BCS und berechne neue Scores

$$SM_{qt}^{QT} = SM^{QT}(RT_{qt}^{QT}) = NWRMSD(RT_{qt}^{QT}) = \frac{1 + e^{-\lambda}}{1 + e^{WRMSD(RT_{qt}^{QT}) - \lambda}} \tag{54}$$

mit

$$WRMSD(RT_{qt}^{QT}) = \sqrt{\frac{1}{RMSS(RT_{qt}^{QT}) \cdot Z(RT_{qt}^{QT})} \sum_{z=1}^{Z(RT_{qt}^{QT})} ((R \cdot \vec{q}_z + \vec{T}) - \vec{t}_z)^2} \tag{55}$$

Die Färbung der einander zugeordneten Residuentupel in Abb. 18 erfolgt nach den $NWRMSD(RT_{qt}^{QT})$-Werten.

2. Normalisiere neue Scores

$$norm(SM_{qt}^{QT}) = \frac{SM_{qt}^{QT}}{max(SM^{QT})} \tag{56}$$

Abb. 19 zeigt die normalisierten Scores. Im Vergleich zu der Scoringmatrix aus der

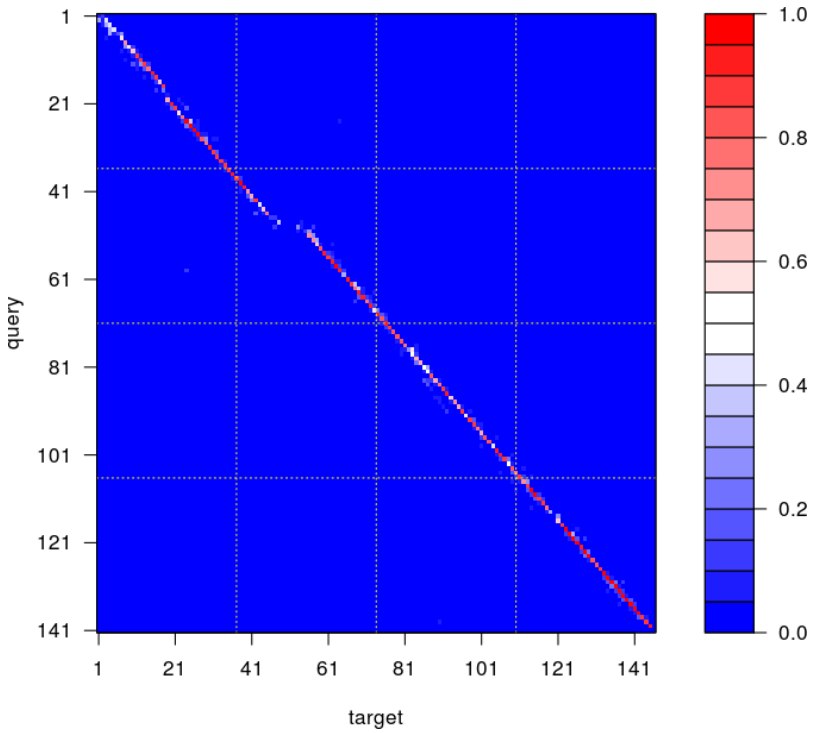


Abbildung 19: Scoringmatrix nach dem BCS-Resampling.

ersten Algorithmusstufe (Abb. 8b), liefert die neue Scoringmatrix eine deutlich bessere Signalstärke.

3. Entferne die Residuentupel in absteigenden Reihenfolge aus der Matrix und blockiere den Zugriff auf die Reihe und Spalte des entfernten Residuentupels nach jedem Entfernen. Auf diese Weise wird eine eineindeutige Residuenzuordnung erreicht. An dieser Stelle findet ein weiterer, benutzerdefinierter Threshold-Parameter Normalized Best Common Substructure Node Score Threshold (NBCSNST)

$$NBCSNST = 0.01$$

seinen Einsatz. Alle Residuentupel mit einer kleineren Ähnlichkeit als 1% zu dem besten Residuentupel werden ignoriert.

4. Die entfernten Residuentupel sind das neue Residuentupel-Set. Das BCS-Resampling liefert 136 richtig-positive von insgesamt 138 Residuentupel. Alle CSs ab der BCS-Ordnung $k = 127$ (Abb. 17) werden mit dem neuen Residuentupel-Set kombinatorisch erweitert. Schritte 1 bis 4 werden wiederholt, solange nach jedem Durchgang eine neue BCS gefunden wird.

COMPS	ATOMS	RMSD	NWRMSD	IDENT	SSIM	CSS	QMCSS	TMCSS	QTMCSS	NBCSS	MCSS
133	859	1.594	0.769	0.459	0.723	0.868	0.819	0.791	0.805	1.0	0.819
139	900	1.80	0.703	0.446	0.707	0.857	0.845	0.816	0.83	0.955	0.845

Tabelle 3: Deskriptoren der BCS und der MCS nach dem BCS-Resampling.

Nach dem BCS-Resampling verschieben sich die BCS (Tab. 3, $COMPS = 133$) und die MCS (Tab. 3, $COMPS = 139$) auf höhere Ordnungen.

2.2.6.5 *Größte gemeinsame Substruktur*

Abb. 20 zeigt die MCS (Tab. 3, $COMPS = 139$) nach dem BCS-Resampling. Deutlich zu

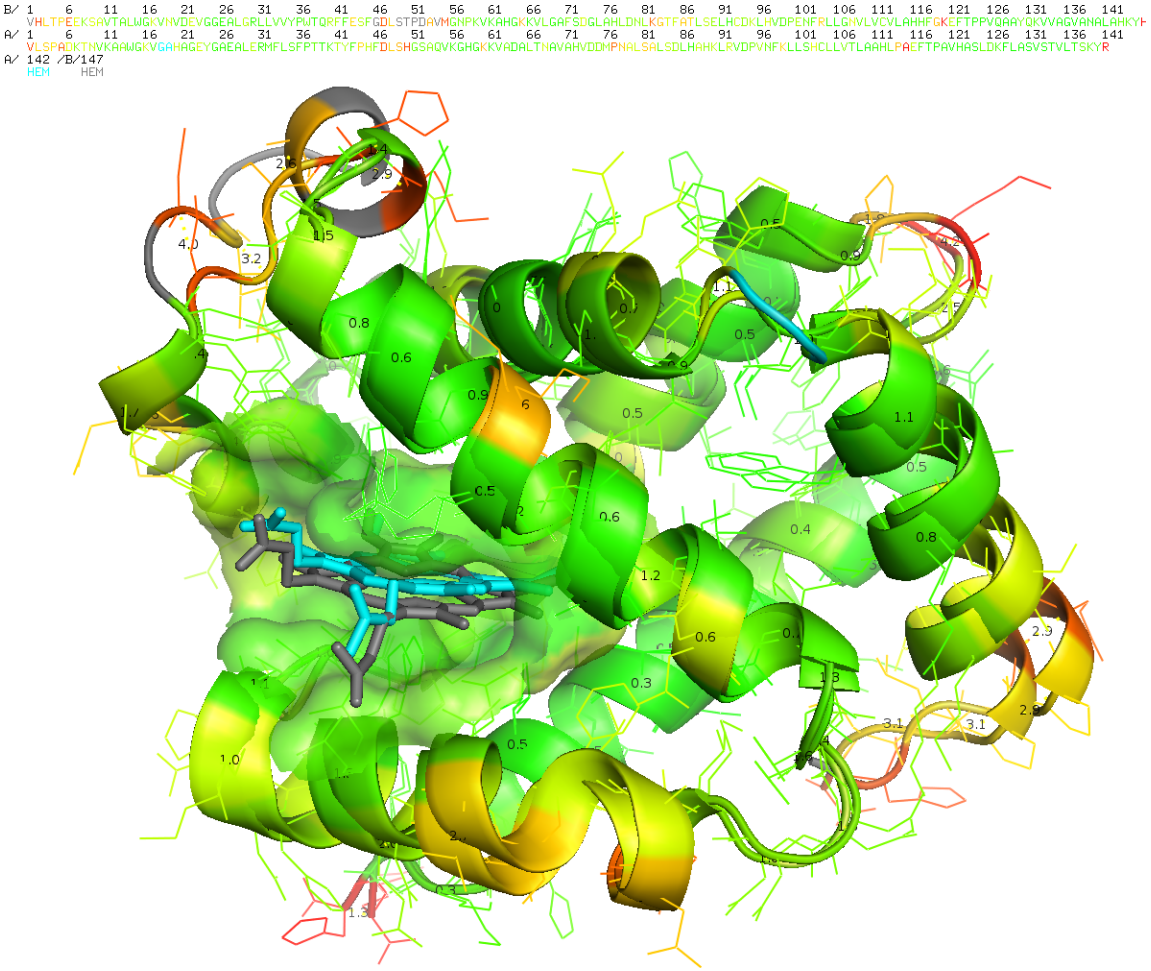


Abbildung 20: Die MCS nach dem BCS-Resampling. Die Färbung der korrespondierenden Residuen erfolgt nach der $NWRMSD(RT_{qt}^{QT})$ (min max) des jeweiligen Residuentupels.

erkennen sind die ergänzten Residuentupel mit größeren konformationellen Unterschieden gegenüber dem Gesamtbild. In dem letzten Algorithmusschritt, dem MCS-Resampling, wird nochmals nach den alternativen Residuentupel und somit nach weiteren alternativen Alignmentkombinationen gesucht:

1. Transformiere alle Residuentupel RT_{qt}^{QT} aus der Scoringmatrix SM^{QT} anhand der Transformationsdaten der MCS und berechne neue Scores SM_{qt}^{QT} analog zum ersten Schritt aus BCS-Resampling. Die Färbung der einander zugeordneten Residuentupel in Abb. 20 erfolgt nach $NWRMSD(RT_{qt}^{QT})$-Werten.

2. Berechne für jede Query-Reihe (Abb. 21) das $max(SM_q^{QT})$ und normalisiere jede Query-Reihe gegen ihren Maximum

$$norm(SM_{qt}^{QT}) = \frac{SM_{qt}^{QT}}{max(SM_q^{QT})} \tag{57}$$

Abb. 21 zeigt die nach den Maximalwerten der Reihen normalisierte Scores.

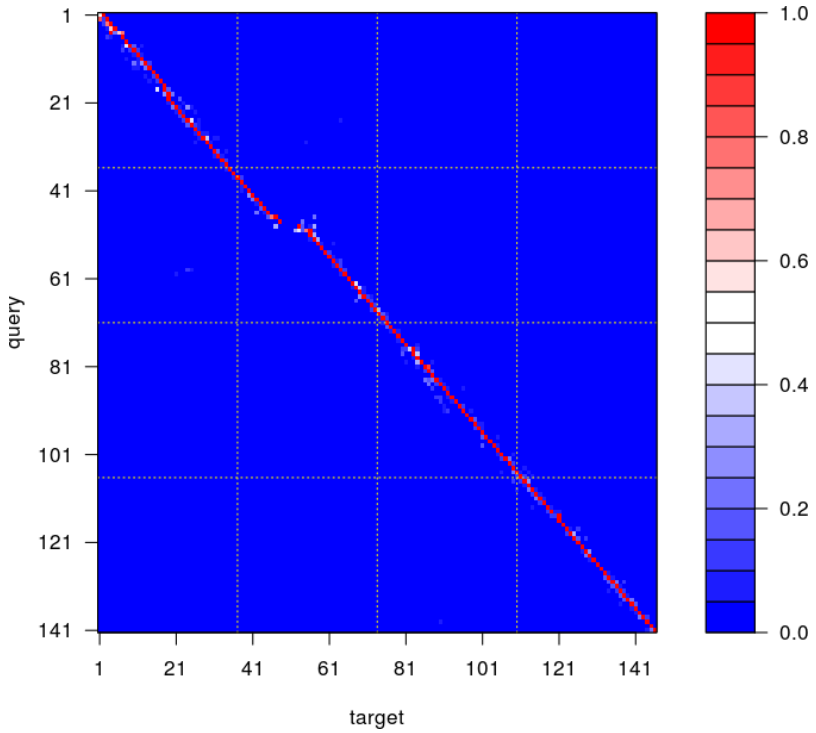


Abbildung 21: Scoringmatrix nach dem MCS-Resampling.

3. Berücksichtige eine Reihe nur dann, wenn

$$max(SM_q^{QT}) \geqq NBCSNST$$

4. Finde alle Zuordnungen mit den normalisierten Scores

$$norm(SM_{qt}^{QT}) \geqq NMCSNST$$

größer oder gleich Normalized Maximum Common Substructure Node Score Threshold (NMCSNST). Das benutzerdefinierte Threshold-Parameter $NMCSNST = 0.5$ sortiert alle Residuentupel mit einer kleineren Ähnlichkeit als 50% zu dem besten Residuentupel einer query-Reihe aus. Im Gegensatz zum BCS-Resampling ist das neue Residuentupel-Set nicht eineindeutig - einem Query-Residuum können mehrere Target-Residuen und einem Target-Residuum können mehrere Query-Residuen entsprechen.

5. Das MCS-Resampling liefert 139 richtig-positive von insgesamt 149 Residuentupel. Alle CSs ab der BCS-Ordnung $k = 1$ werden mit dem neuen Residuentupel-Set kombinatorisch erweitert.

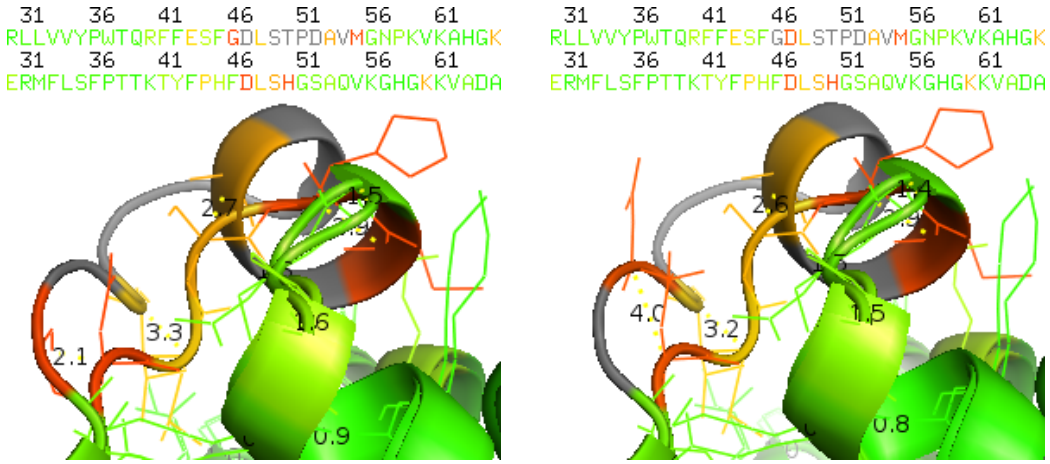
COMPS	ATOMS	RMSD	NWRMSD	IDENT	SSIM	CSS	QMCSS	TMCSS	QTMCSS	NBCSS	MCSS
134	862	1.597	0.765	0.447	0.714	0.87	0.827	0.798	0.812	1.0	0.827
139	897	1.773	0.705	0.439	0.702	0.857	0.845	0.816	0.83	0.956	0.845
139	900	1.80	0.703	0.446	0.707	0.857	0.845	0.816	0.83	0.953	0.845

Tabelle 4: Deskriptoren der BCS und der MCS nach dem MCS-Resampling.

Das MCS-Resampling resultiert in einer neuen BCS (Tab. 4, $COMPS = 134$) und zwei MCSs (Tab. 4, $COMPS = 139$). Der einzige Unterschied zwischen den beiden MCSs sind dem Query-Residuum R^Q_{ASP47} zugeordneten Target-Residuen R^T_{GLY46} (atoms (ATOMS) $= 897$, Abb. 22a) und R^T_{ASP47} ($ATOMS = 900$, Abb. 22b). Der Unterschied in der Anzahl der Atome entsteht aus der Differenz $Z(R^T_{ASP47}) - Z(R^T_{GLY46}) = 3$ (Abb. 5). Währen die MCS mit weniger Atomen über eine bessere $RMSD = 1.773\mathring{A}$ verfügt, verfügt die MCS mit mehr Atomen über eine höhere Substitutionsähnlichkeit $SSIM = 0.707$. Die MCSSs der beiden Alignmentkombinationen unterscheiden sich nur geringfügig

$$MCSS(CS^{139,900}) - MCSS(CS^{139,897}) = 0.84502333 - 0.8450054 = 1.793 \cdot 10^{-5}$$

wobei die größere und gleichzeitig physiko-chemisch bessere MCS etwas höher bewertet ist. Beide alternativen MCSs zeugen von einer feinen Balance zwischen den geometrischen und



(a) Geometrisch bessere MCS. Der kleinere RMSD-Wert auf der rechten roten Linie in der Abbildung Abb. 23.

(b) Physiko-chemisch bessere MCS. Der größere RMSD-Wert auf der rechten roten Linie in der Abbildung Abb. 23.

Abbildung 22: MCS vs. MCS.

physiko-chemischen Eigenschaften. Die Gesamtheit aller gefundenen und gespeicherten CSs zweier Strukturen wird im Rahmen von EpitopeMatch als ein *Match* bezeichnet, die jeweilige CS eines Match als *Permutation*. Abb. 23 zeigt Deskriptorwerte aller 1176 gespeicherten Permutationen. Die graue senkrechte Linie markiert die BCS. Die roten senkrechten Linien schließen den Bereich ein, in dem

$$BCSS(CS) \geqq 0.95$$

die CSs mindestens zu 95% der BCS ähneln. In diesem Fall handelt es sich um 107 Alternativalignments, die inklusive BCS und MCS interessant sein könnten. Der höchste MCSS

$$MCSS(CS^{139,900}) = 0.84502333$$

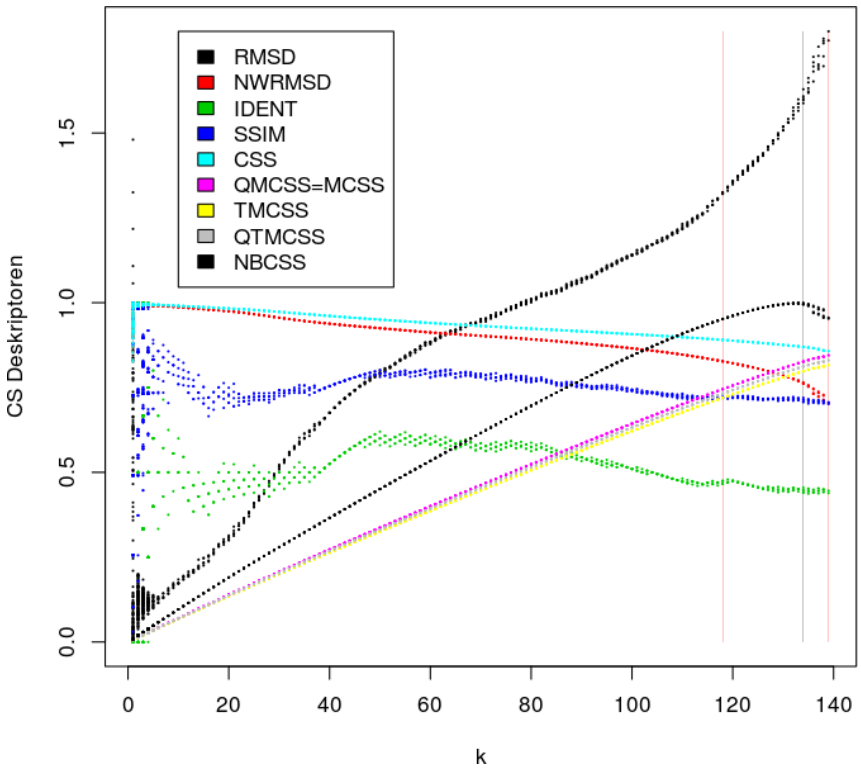


Abbildung 23: Deskriptoren der finalen CSs. Die graue Linie markiert die BCS. Die roten Linien markieren den Bereich mit den zu der BCS sehr ähnlichen CSs.

sagt aus, dass unter der Berücksichtigung der geometrischen (Abb. 5) und physiko-chemischen (Abb. 7) Eigenschaften die Ketten A und B aus $2DN2$ zu 84.5% einander ähnlich sind. Je nach Wunsch des Benutzers setzt der Algorithmus mit der Suche nach weiteren Matches (Epitopen) fort. Dabei kehrt der Algorithmus zu dem Schritt *initiales Alignment* (Abs. 2.2.5) zurück und sucht nach dem nächsten Match unter Ausschluss der Residuentupel aus den bislang gefundenen MCSs. Für das *kombinatorische Resampling* (Abs. 2.2.6) des initialen Alignments der Ketten A und B aus $2DN2$ braucht der Algorithmus $\approx 1800ms$, sodass beide Stufen insgesamt $\approx 500ms + 1800ms = 2300ms$ in Anspruch nehmen.

2.2.7 *Induced-Fit & Hinge-Bending*

Die erste Algorithmusstufe (Abs. 2.2.5) liefert die initiale Alignmentkombination $K^A \in A$ mit $min(|Q|, |T|)$ Residuentupel RT^{QT}. Die zweite Algorithmusstufe (Abs. 2.2.6) leitet von der initialen Alignmentkombination unter Einsatz der Deskriptoren (Abs. 2.2.6.1) ein Match ab. Ein *Match* ist die Menge aller gemeinsamen Substrukturen CS (Abs. 2.2.6.2), die als alternativen Alignmentkombinationen um den Kern der Ähnlichkeit, die beste gemeinsame Substruktur BCS (Abs. 2.2.6.4), verteilt sind. Die größte gemeinsame Substruktur MCS (Abs. 2.2.6.5) gehört ebenfalls zu dieser Menge. Wie groß diese Menge ist, d.h., wie viele alternativen Alignmentkombinationen pro Match ermittelt werden, bestimmen die gewählten Threshold-Parameter, die sowohl in der ersten als auch in der zweiten Algorithmusstufe ihren Einsatz finden. Für die Ermittlung des gesamten Strukturalignments ist ein einziges Match oft nicht ausreichend. Dies ist insbesondere dann der Fall, wenn die *Apo*-Struktur (ungebundener Zustand) während der Wechselwirkung mit einem Liganden durch die induzierte Anpassung, das sogenannte Induced-Fit [95], ihre Konformation ändert. *Apo*-Struktur

in Kombination mit einem oder mehreren Liganden bezeichnet man als *Holo*-Struktur (gebundener Zustand). Je dramatischer die relative Konformationsänderung der *Apo*- bzw. *Holo*-Struktur, desto höher ist die Wahrscheinlichkeit, dass die Suche nach ihrer Ähnlichkeit aus dem Rahmen der üblichen Threshold-Parameter fällt. EpitopeMatch sieht die Möglichkeit vor, nach $N > 1$ Matches zu suchen, um diese in der dritten Algorithmusstufe zu einem gesamten Strukturalignment miteinander zu kombinieren. Die folgenden Abschnitte beschreiben die dritte und zugleich die letzte Algorithmusstufe und behandeln ausführlich ein *Holo-Apo*-Strukturpaar, mit dem die Arbeit an EpitopeMatch begann.

2.2.7.1 *CSs im Auge des Betrachters*

Abb. 24 zeigt die *Apo*-Struktur 1OMP (Target) eines an dem aktiven Transport und Chemotaxis beteiligten, Maltodextrin bindenden Proteins. Diese Struktur ist überlagert mit der an

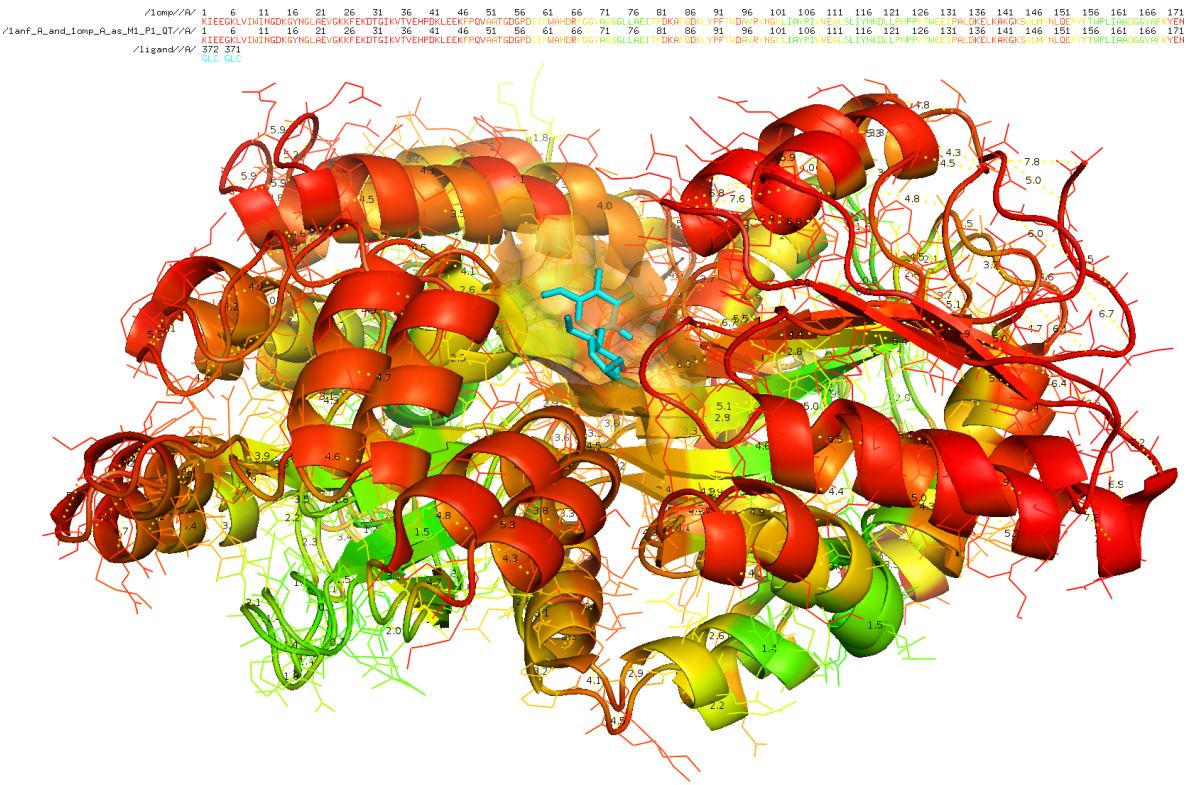


Abbildung 24: 1ANF & 1OMP, vollständig überlagert. Die Färbung der korrespondierenden Residuen erfolgt nach der $NWRMSD(RT_{qt}^{QT})$ (min max) des jeweiligen Residuentupels.

die Maltose gebundenen *Holo*-Struktur 1ANF (Query) desselben Proteins. Die *Apo*-Struktur gilt als ein kristallografischer Nachweis für ein großes Hinge-Twist bzw. Hinge-Bending (eine gelenkartige Verdrehung bzw. Krümmung), das durch die Ligand-Bindung induziert wird und zu einer relativen Positionsänderung der beiden globulären Domänen führt [152]. Beide Strukturen (Proteinanteil) sind jeweils 370 Aminosäuren bzw. 2585 Atome (ohne Wasserstoffe) groß. Die einander entsprechenden Residuentupel sind nach der NWRMSD (54) gefärbt, die nach der Überlagerung der beiden gesamten Strukturen für jedes Residuentupel einzeln berechnet wird. *Rot* bedeutet, dass: die Residuen eines Residuentupels nach der Überlagerung weit voneinander entfernt sind; ihre Konformation sich ggf. stark unterscheidet; ihre physiko-chemischen Eigenschaften sich stark unterscheiden (da die Identität der beiden Strukturen 100% ist, gibt es in diesem Fall keine physiko-chemischen Unterschiede). *Grün* bedeutet das Gegenteil von *rot*. Das Farbspektrum verläuft von *grün* nach *rot* über

gelb. *Gelb* markiert die mittlere Residuenähnlichkeit. Die Maltose aus dem Komplex 1ANF ist in *cyan* dargestellt und liegt in einer Furche zwischen den beiden globulären Domänen begraben. Die gesuchte Querystruktur (QS) ist vollständig erkannt und ist somit die größte gemeinsame Substruktur (MCS) der Targetstruktur (TS). Obwohl die MCS bekannt ist, liefert die subjektive Betrachtung der beiden Strukturen nach der Überlagerung (Abb. 24) keine genaue Aussage über die mögliche Aufteilung der Struktur in einzelne Domänen. Der relativ hohe Wert für die Rigidität (RGD) mit $RGD = CSS = MCSS = 0.664 \triangleq 66.4\%$ ($FLX = 1.0 - 0.664 = 0.336 \triangleq 33.6\%$) bei einer relativ niedrigen $RMSD = 3.939\text{Å}$ (AllAtoms, Abb. 5), am oberen Rand der Zwielichtzone (Abb. 10), zeugt von dem Vorhandensein mindestens einer großen rigiden gemeinsamen Substruktur. Abb. 25 zeigt ins-

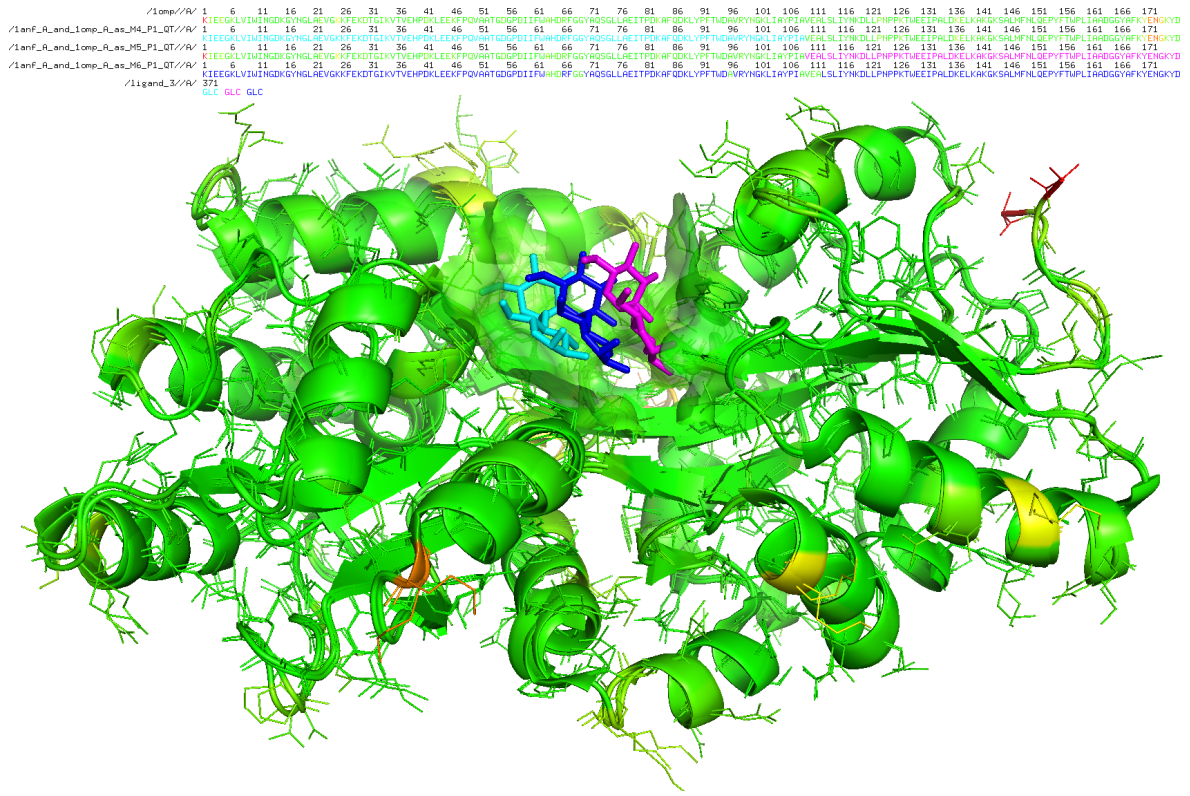


Abbildung 25: 1ANF & 1OMP, domänenweise überlagert. Die Lage der Maltose entspricht der Überlagerung der jeweiligen Domäne (links, rechts) bzw. der Hinge-Bending-Region (zentral).

gesamt drei Überlagerungen. Die MCS ist in zwei gemeinsame Substrukturen (CSs) zerlegt, die jeweils eine Domäne (links (*cyan*), rechts (*magenta*)) repräsentieren und gemeinsam die vollständige Struktur abdecken. Die dritte CS ist die zentral gelegene Hinge-Bending-Region, die eine Teil- und Schnittmenge der Residuentupel der beiden Domänen ist. Die Interpretation des Farbspektrums ist analog zu der Abb. 24. Die dreifach dargestellte Maltose aus dem Komplex 1ANF resultiert aus der Transformation der Domäne 1 (*cyan*), Domäne 2 (*magenta*) und der Hinge-Bending-Region (*blau*) auf das Target. Aufgrund der verhältnismäßig kleinen $DRMSD = 0.99\text{Å}$ (AllAtoms, Abb. 5, 51) und der hohen Domain RiGiDity (DRGD) $DRGD = DCSS = DMCSS = 0.965 \triangleq 96.5\%$ gelangt man zu der Aussage, dass die Struktur zumindest vor und nach dem Induced-Fit aus zwei stabil gefalteten Domänen besteht. Die Domain FLeXibility (DFLX), d.h. die Flexibilität der Domänen exklusive ihrer Verschiebung, beträgt lediglich $DFLX = 1.0 - 0.965 = 0.035 \triangleq 3.5\%$ und kann auf Bewegungen der Residuenreste und die geringfügigen Konformationsänderungen des Rückgrats

zurück geführt werden. Die letzteren bilden die tatsächliche Hinge-Bending-Region (Abb.

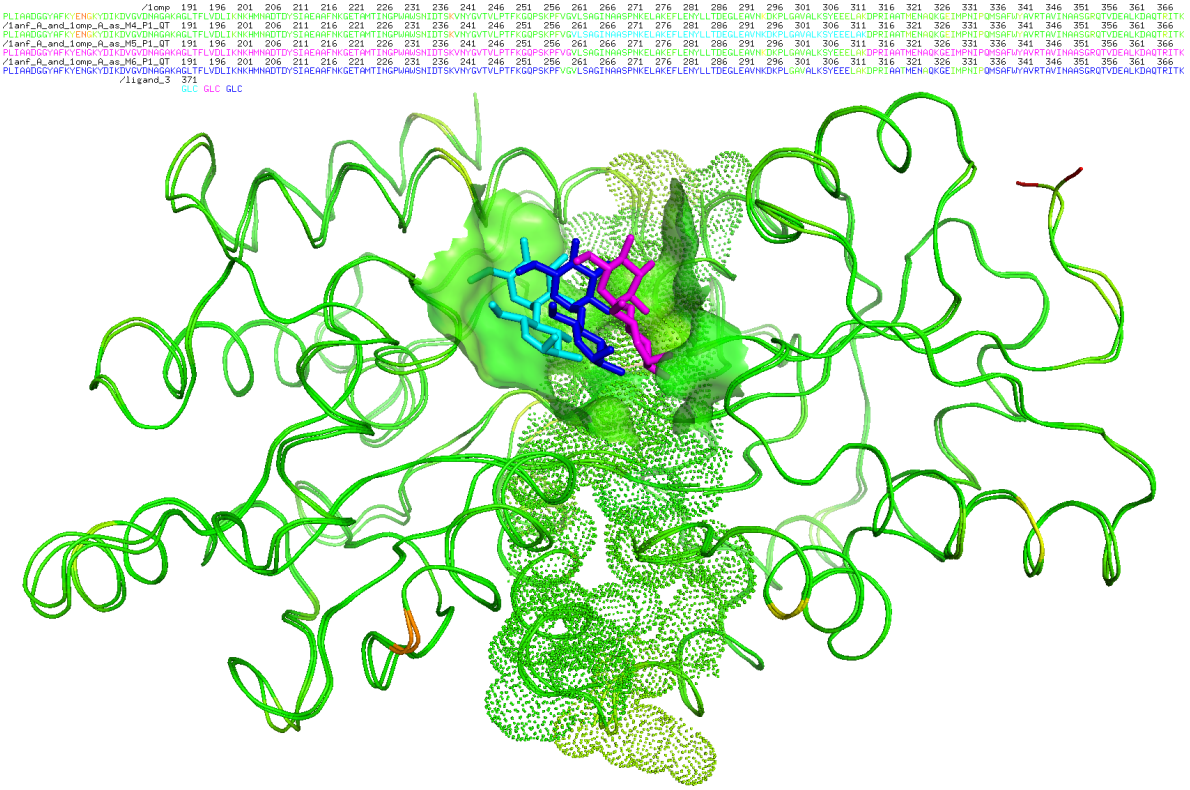


Abbildung 26: 1ANF & 1OMP, Hinge-Bending-Region. Die Residuen der Hinge-Bending-Region sind als Punkte (*dots*) hervorgehoben.

26, *dots*). Zu der Hinge-Bending-Region gehören außer GLU111, VAL261, ALA312 [152] 28 weitere Residuen, ihre unmittelbaren Nachbarn. Die ersten beiden Residuen liegen in den β-Faltblättern und das letztere im Turn zwischen den Helices. Diese drei Rückgratabschnitte verbinden die beiden Domänen. Die β-Faltblätter verlaufen unterhalb der dargestellten Maltose-Molekülen, wobei GLU111 in die 5$\AA$-Umgebung der Maltose hineinragt und somit ein Teil des Maltose-Epitops ist. Es ist deutlich zu erkennen, dass die Konformation der beiden Domänen sowohl im *Apo-*, als auch im *Holo*-Zustand größtenteils konserviert ist. Ihre relative Positionsänderung ist u.a. das Resultat der Torsionswinkeländerung der oben genannten Residuen. Inwiefern EpitopeMatch dies abzuschätzen und zu beurteilen vermag, wird im Folgenden gezeigt.

2.2.7.2 *CSs aus der Sicht von* EpitopeMatch

Im Verlauf des kombinatorischen Resampling (Abs. 2.2.6) werden die einzelnen Residuentupel RT_{qt}^{QT} zu den gemeinsamen Substrukturen CSs aufgebaut. Pro Ordnung k (Abb. 15) werden nur die am höchsten bewerteten CSs behalten. Durch das Hinzufügen der neuen Residuentupel zu der jeweiligen CS der Ordnung k werden die CSs der nächsthöheren Ordnung $k+1$ erzeugt, usw... Auf diese Weise werden die Residuentupel der im Rahmen der Threshold-Parameter bleibenden CSs immer wieder für den Aufbau neuer CSs verwendet. Die meisten berechneten CSs (außerhalb der Threshold-Parameter) werden verworfen, weil ihre Speicherung und Weiterführung sonst den Gegenstand der kombinatorischen Explosion ausmachen würden. Ein Teil der mit ihrer Berechnung gewonnenen Information wird jedoch genutzt. Unabhängig davon, ob eine CS behalten oder verworfen wird, werden ihre

Transformationsdaten und der MCSS (48) für die Bewertung der an ihrem Ausbau beteiligten Residuentupel verwendet. Pro Residuentupel wird ein Residuum Tuple Score (RTS)

$$RTS(RT_{qt}^{QT}) = NWRMSD(RT_{qt}^{QT}) \cdot MCSS(P(K^A)), RT_{qt}^{QT} \in K^A \tag{58}$$

berechnet. Der RTS bewertet die geometrische und die physiko-chemische Ähnlichkeit NWRMSD (54) eines Residuentupels im Bezug auf die größenabhängige Ähnlichkeit MCSS (48) der gemeinsamen Substruktur. Pro Match m wird eine Scoring-Summenmatrix

$$SSM_m^{QT}(RT_{qtm}^{QT}) = SSM_m^{QT}(RT_{qtm}^{QT}) + RTS(CS(RT_{qtm}^{QT})) \tag{59}$$

erzeugt, in der alle berechneten RTSs gesammelt werden. Residuentupel, die am häufigsten in den CSs eines Matchs vorkommen, werden auf diese Weise am höchsten bewertet. Tab. 5 fasst die Deskriptoren der MCSs und BCSs der beiden gefundenen Matches zusammen.

TYPE	MATCH	COMPS	ATOMS	RMSD	NWRMSD	IDENT	SSIM	CSS	NBCSS	MCSS
MCS	1	288	1963	2.028	0.662	0.809	0.93	0.887	0.936	0.69
MCS	2	264	1789	2.308	0.573	0.72	0.894	0.846	0.882	0.603
BCS	1	250	1698	1.437	0.814	0.888	0.959	0.935	1.0	0.632
BCS	2	215	1449	1.494	0.798	0.851	0.944	0.926	1.0	0.538

Tabelle 5: 1ANF & 1OMP, MCSs und BCSs der Matches 1 & 2.

Abb. 27 zeigt den Verlauf der Deskriptoren der Ordnungen $1 < k < 288$ des größeren

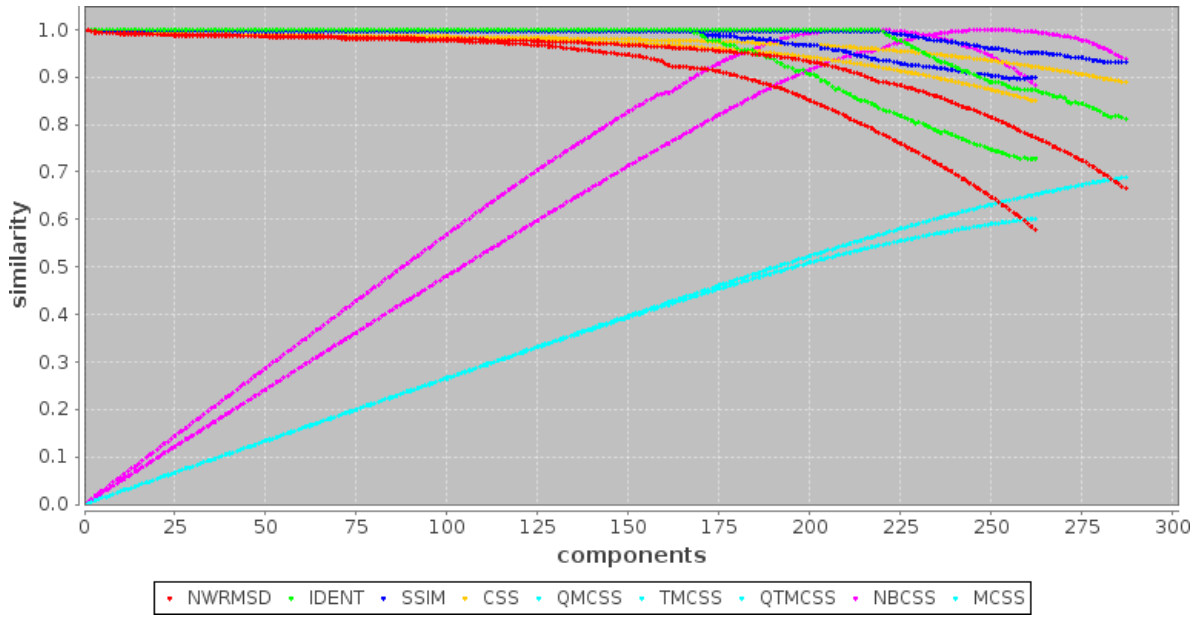


Abbildung 27: 1ANF & 1OMP, Deskriptoren des Match 1 (alle Kurven verlaufen bis $k = 288$) & des Match 2 (alle Kurven verlaufen bis $k = 264$). Die Identität (grün) ist bis $k = 219$ des Match 1 und bis $k = 167$ des Match 2 gleich 1.0. Es handelt sich hierbei um die beiden Domänen. Die CSs der höheren Ordnungen sind falsch-positiv.

Match 1 und der Ordnungen $1 < k < 264$ des kleineren Match 2. Da die Primärstrukturen von 1ANF und 1OMP identisch sind, gilt für das jeweilige Match $MCSS = QMCSS = TMCSS = QTMCSS$. Als Substitutionsmatrix wurde die sigmoid normalisierte $BM_{sigmoid}^{62}$-Matrix (16) gewählt. Während der Suche nach der BCS (Abs. 2.2.6.4) bleibt die Identität bis $k = 219$ des Match 1 und bis $k = 167$ des Match 2 gleich 1.0. In den höheren Ordnungen

greift das geometrische Rauschen. Die Residuentupel mit identischen Aminosäuren ergeben in der Summe $219 + 167 = 386 > 370$. Die Vermutung liegt also nahe, dass die beiden Domänen vollständig erkannt sind. Anhand von NBCSS kann man entscheiden, dass die gemeinsamen Substrukturen, die kleiner oder gleich als die BCSs der Ordnungen $k \leqq 250$ (Match 1, $k = 250, NBCSS = 1.0$) bzw. $k \leqq 215$ (Match 2, $k = 215, NBCSS = 1.0$) sind, zum Kern der Ähnlichkeit gehören.

Für die Bestimmung des Gesamtstrukturalignments eignet sich der Vergleich der Scoring-Summenmatrizen (59). Am Verlauf der Deskriptoren (Abb. 27) ist zu erkennen, das die größere Domäne bezüglich der Ähnlichkeit dominiert. Offensichtlich ist sie stabiler als die kleinere Domäne und konserviert ihre Konformation im Verlauf von Induced-Fit besser als die kleinere Domäne. Aus diesem Grund wird sie als erstes erkannt. Abb. 28 zeigt die

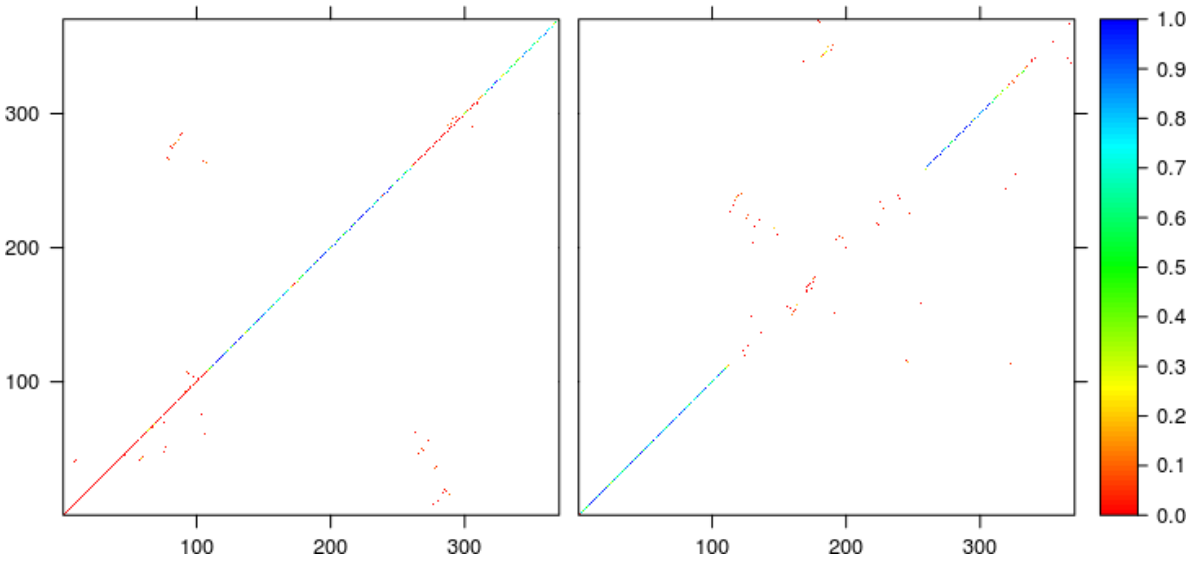


Abbildung 28: Normalisierten Scores der Matches 1 & 2.

normalisierten RTS-Summen (59) der beiden Matches. Die Normalisierung der Scores ist lediglich für die Darstellung erforderlich. Während die RTS-Summen der richtig-positiven Residuentupel für Match 1 (links) alle vorhanden sind (Diagonale), fehlen alle Residuentupel im Match 2 (rechts, Diagonale), die zu der BCS des Match 1 gehören. Mit anderen Worten, wird aus der initialen Alignmentkombination eine BCS ermittelt, so werden ihre Residuentupel aus der Menge der initialen Alignmentkombination gelöscht, wobei der Rest für die Suche nach der weiteren BCS verwendet wird. Diese Vorgehensweise wiederholt sich bis der Vorrat an Residuentupeln der initialen Alignmentkombination erschöpft ist.

Wenn beim Vergleich zweier Strukturen mehrere MCSs erkannt werden, und die entsprechenden BCSs sich mit mindestens einem Residuentupel überschneiden, dann liegen vermutlich mehrere Domänen vor. Hierbei muss angemerkt werden, dass die einzelnen Domänen erst dann sichtbar werden, wenn eine spürbare Domänenverschiebung (Query gegenüber Target) vorliegt. Die Domänenzugehörigkeit der Residuentupel kann in diesem Fall anhand der Überlagerung der beiden Matrizen

$$RT_{qt}^{QT} \in \begin{cases} m & SSM_m^{QT}(RT_{qtm}^{QT}) \geq SSM_n^{QT}(RT_{qtn}^{QT}) \\ n & sonst \end{cases} \tag{60}$$

bestimmt werden. Aus diesen Daten werden vier neue Alignmentkombinationen bzw CSs ermittelt (Tab. 6): die größere Domäne 1 $MCS_{d(1)}$ (Abb. 25, links), 208 Residuentupel; die

TYPE	MATCH	DCOMPS	DATOMS	DRMSD	DNWRMSD	DIDENT	DSSIM	DCSS	DNBCSS	DMCSS
MCS	d(1)	208	1457	1.011	0.902	1.0	0.998	0.964	0.995	0.542
MCS	d(2)	162	1128	0.968	0.908	1.0	0.998	0.967	0.988	0.423
MCS	h(d(1), d(2))	31	194	1.168	0.879	1.0	0.997	0.93	1.0	0.078
MCS	c(d(1), d(2))	370	2585	3.939	0.243	1.0	0.998	0.664	0.46	0.664
DMCS	d(1)+d(2)	**370**	**2585**	**0.992**	**0.905**	**1.0**	**0.998**	**0.965**	**0.992**	**0.965**

Tabelle 6: 1ANF & 1OMP, MCSs der kombinierten Matches 1 & 2.

kleinere Domäne 2 $MCS_{d(2)}$ (Abb. 25, rechts), 162 Residuentupel; die Hinge-Bending-Region $MCS_{h(d(1),d(2))}$ (Abb. 26, *dots*), 31 Residuentupel; und das vollständige Strukturalignment $MCS_{c(d(1),d(2))}$ (Abb. 24), 370 Residuentupel. Tab. 6 enthält zusätzlich die Domain Score (DS)s (fett), mit deren Hilfe Rückschlüsse auf die Rigidität bzw. Flexibilität der Substrukturen gezogen werden können. Alle vier Matches bestehen ausschließlich aus richtig-positiven Residuentupeln ($IDENT = 1.0$). Abb. 29 zeigt den Verlauf der Deskriptoren der Ordnungen

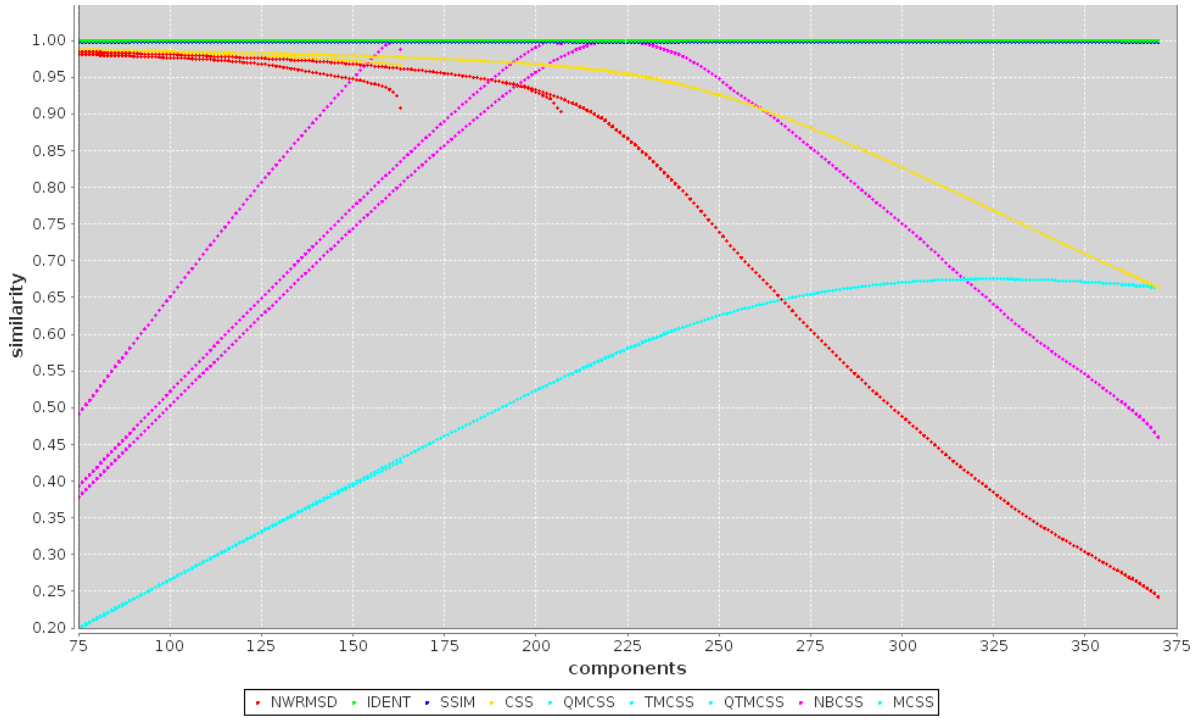


Abbildung 29: 1ANF & 1OMP, Deskriptoren des Match 1 (alle Kurven bis $k = 208$), des Match 2 (alle Kurven bis $k = 162$) und ihre Kombination (alle Kurven bis $k = 370$). Der Scheitelpunkt der NBCSS-Parabel der Kombination (magenta) zeigt an, dass die größere Domäne (Match 1) die BCS der beiden Strukturen ist.

$75 < k < 208$ der Domäne 1 $MCS_{d(1)}$, der Ordnungen $75 < k < 162$ der Domäne 2 $MCS_{d(2)}$ und der Ordnungen $75 < k < 370$ des vollständigen Strukturalignments $MCS_{c(d(1),d(2))}$. Da das Match $MCS_{c(d(1),d(2))}$ aus den Residuentupeln der Matches $MCS_{d(1)}$ und $MCS_{d(2)}$ besteht, können seine größenunabhängigen Deskriptoren (RMSD, NWRMSD, IDENT, SSIM, CSS, NBCSS) und größenabhängigen Deskriptoren (COMPS, ATOMS, QMCSS, TMCSS, QTMCSS, MCSS) in die Domänen-Form gebracht werden (51, 52). Auf diese Weise können z.B. die DRMSD und der DMCSS wie folgt berechnet werden

$$DRMSD(MCS_{d(1)}, MCS_{d(2)}) = \frac{208 \cdot 1.011\mathring{A} + 162 \cdot 0.968\mathring{A}}{370} = 0.992\mathring{A}$$

$$DMCSS(MCS_{d(1)}, MCS_{d(1)}) = 0.542 + 0.423 = 0.965 \triangleq 96.5\%$$

Das Induced-Fit setzt sich demnach aus zwei wesentlichen Komponenten zusammen, aus der relativen Konformationsänderung der Residuen der beiden Domänen und aus der relativen Verschiebung der beiden Domänen

$$RMSD(MCS_{c(d(1),d(2))}) - DRMSD(MCS_{d(1)}, MCS_{d(2)}) = 3.939\mathring{A} - 0.992\mathring{A} = 2.947\mathring{A}$$

Schließt man die Domänenverschiebung aus, so stellt man fest, dass die beiden Domänen im Verlauf von Induced-Fit ihre native Konformation beibehalten. Die Hinge-Bending-Region ist die Schnittmenge aus den BCSs der beiden Matches (Tab. 5)

$$hinge \triangleq BCS(MATCH_1) \cap BCS(MATCH_2) \tag{61}$$

Tab. 7 zeigt das aus dem Strukturalignment resultierende Sequenzalignment. Die Sequenz

```
  1        10        20        30        40        50        60        70        80
  KIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEIT
1 ------------------------------------------------------------------------------
2 KIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEIT
h ------------------------------------------------------------AHD--GG-----------
  81       90        100       110       120       130       140       150       160
  PDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPL
1 ----------------------------VEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPL
2 PDKAFQDKLYPFTWDAVRYNGKLIAYPIA-----------------------------------------------
h --------------A-----------AVEA-----------------------------------------------
  161      170       180       190       200       210       220       230       240
  IAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKV
1 IAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKV
2 ------------------------------------------------------------------------------
h ------------------------------------------------------------------------------
  241      250       260       270       280       290       300       310       320
  NYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAAT
1 NYGVTVLPTFKGQPSKPFVG-----------------------------------------------DPRIAAT
2 --------------------VLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAK-------
h ------------------VGV-------------------------------------GAV--------LAKDPRI--T
  321      330       340       350       360       370
  MENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDAQTRITK
1 MENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDAQTRITK
2 -------------------------------------------------
h ---A----TMPNIP-----------------------------------
```

Tabelle 7: 1ANF & 1OMP, Domänen 1, 2 & hinge, FASTA.

der 1ANF ist auf zwei Reihen verteilt. Jede Reihe symbolisiert das entsprechende Domäne. Die alignierten Sequenzabschnitte der 1ANF wechseln sich insgesamt 3 Mal ab. Und zwar genau dort, wo sich die drei zentralen Residuen GLU111, VAL261 und ALA312 mit den größten Torsionswinkeländerungen [152] befinden. Diese Residuen liegen entweder als Grenzresiduen der Domänen-Schnittstellen (im Fall von VAL261) oder als ihr unmittelbarer Nachbar (im Fall von GLU111 und ALA312) vor. Demnach gibt es exakt 3 Hinge-Bending-

Regionen, die die beiden Domänen miteinander verbinden. Die 31 *hinge*-Residuen (Abb. 26, *dots*) sind in der dritten Reihe (h) der Tabelle enthalten. Die Bewertung der *hinge*-Residuen

$$HS(RT_{qt}^{QT}) = \frac{HSQ(RT_{qt}^{QT}) = \begin{cases} \frac{SSM_n^{QT}(RT_{qtn}^{QT})}{SSM_m^{QT}(RT_{qtm}^{QT})} & SSM_m^{QT}(RT_{qtm}^{QT}) \geq SSM_n^{QT}(RT_{qtn}^{QT}) \\ \frac{SSM_m^{QT}(RT_{qtm}^{QT})}{SSM_n^{QT}(RT_{qtn}^{QT})} & sonst \end{cases}}{HS_{max}} \tag{62}$$

ist mittels Hinge Score Threshold (HST) $0.0 \leq HST \leq 1.0$ regulierbar. Je ähnlicher die RTSs der gemeinsamen Residuentupel zweier Domänen sind, desto höher ist der Hinge Score (HS). Die 31 *hinge*-Residuentupel aus diesem Beispiel sind die vollständige Schnittmenge der beiden BCSs, d.h. $HST = 0.0$. Die höheren HST-Werte würden die ausgegeben Menge der *hinge*-Residuentupel auf die Residuen GLU111, VAL261, ALA312 und ihre unmittelbaren Nachbarn reduzieren. Die verhältnismäßig niedrigen Deskriptorenwerte der Hinge-Bending-Region $MCS_{h(d(1),d(2))}$ (Tab. 6) zeugen von einer spürbaren Torsionswinkeländerung der entsprechenden Residuen.

EpitopeMatch ist hiermit in der Lage: die einzelnen Domänen zu erkennen; diese miteinander zu kombinieren; eine Aussage über die Konformationsänderung der Domänen und ihre relative Verschiebung zu treffen; und die Residuen der Hinge-Bending-Regionen zu lokalisieren.

2.2.8 *Diskussion*

Insgesamt verfügt EpitopeMatch über die folgenden Eigenschaften:

1. Der Algorithmus terminiert im Rahmen der gesetzten Threshold-Parameter. Die Threshold-Parameter definieren einen Korridor, in dem die Alignmentkombinationen gebildet und ausgewertet werden.
2. Die Laufzeit und die Qualität ist steuerbar:
   a) Je enger der Korridor, desto weniger Kombinationsfreiheit. Vorteil: schneller. Nachteil: weniger alternativen Alignmentkombinationen.
   b) Je weiter der Korridor, desto mehr Kombinationsfreiheit. Vorteil: mehr alternative Alignmentkombinationen. Nachteil: langsamer.
3. Die Bandbreite ist groß. Anwendbar auf beliebige Biopolymere: verzweigt oder unverzweigt; und ihre Substrukturen: kontinuierlich oder diskontinuierlich.
4. Behandlung von Strukturen der typischen Domänengröße von $\approx$ 150 Aminosäuren innerhalb von $\approx 2s$ (auf einem Kern des Xeon X5650) ist akzeptabel. Die Suche nach einzelnen Strukturen in der gesamten Sturkturdatenbank auf handelsüblichen PCs durchführbar. Hochdurchsatz-Kreuzvergleiche (PDB vs. PDB) auf einem Cluster-System möglich.
5. Erzeugte Datenmenge verwendbar für:
   a) Struktur- und vor allem Substruktur-Klassifizierung
   b) Vorhersage der Funktionalität
   c) Strukturmodellierung durch Residuentransplantation

2.3 ANALYSE

In diesem Abschnitt präsentierten Beispiele konzentrieren sich auf die Qualität, Performance und Bandbreite von EpitopeMatch.

2.3.1 *Proteine*

Proteine sind der Hauptanwendungsbereich von EpitopeMatch. In diesem Abschnitt wird auf die Resultate der ersten Version von EpitopeMatch Jakuschev and Hoffmann [77] zurückgegriffen. Konzeptuell unterscheiden sich die beiden Versionen nicht. Algorithmisch hat sich die neue Version von dem graphentheoretischen Ansatz vollständig getrennt, wobei dieser durch einen kombinatorischen ersetzt worden ist.

2.3.1.1 *Homologe Strukturen*

Im ersten Anwendungsbeispiel der Publikation werden die homologen Ketten *A* und *B* aus 2DN2 [129] miteinander verglichen. Der Vergleich der beiden Strukturen begleitet die Veranschaulichung des Algorithmus in Abs. 2.2.5 und Abs. 2.2.6. Das gewählte Backbone(N,Cα,C′)-Template (Abb. 30) berücksichtigt die interatomaren Distanzen zwischen den

components
20 / 20 / 14810

ALA
ARG
ASN
ASP
CYS
GLN
GLU
GLY
HIS
ILE
LEU
LYS
MET
PHE
PRO
SER
THR
TRP
TYR
VAL

components and their atoms

	0 / 20	0 / 387		20 / 20		60 / 387								
20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ALA 1	N	CA	C	O	CB	OXT	H	H2	HA	HB1	HB2	HB3	HXT		
ALA 1	1	2	3												
ARG 1	N	CA	C	O	CB	CG	CD	NE	CZ	NH1	NH2	OXT	H	H2	HA
ARG 1	1	2	3												
ASN 1	N	CA	C	O	CB	CG	OD1	ND2	OXT	H	H2	HA	HB2	HB3	HD21
ASN 1	1	2	3												
ASP 1	N	CA	C	O	CB	CG	OD1	OD2	OXT	H	H2	HA	HB2	HB3	HD2
ASP 1	1	2	3												
CYS 1	N	CA	C	O	CB	SG	OXT	H	H2	HA	HB2	HB3	HG	HXT	
CYS 1	1	2	3												
GLN 1	N	CA	C	O	CB	CG	CD	OE1	NE2	OXT	H	H2	HA	HB2	HB3
GLN 1	1	2	3												
GLU 1	N	CA	C	O	CB	CG	CD	OE1	OE2	OXT	H	H2	HA	HB2	HB3
GLU 1	1	2	3												
GLY 1	N	CA	C	O	OXT	H	H2	HA2	HA3	HXT					
GLY 1	1	2	3												
HIS 1	N	CA	C	O	CB	CG	ND1	CD2	CE1	NE2	OXT	H	H2	HA	HB2
HIS 1	1	2	3												
ILE 1	N	CA	C	O	CB	CG1	CG2	CD1	OXT	H	H2	HA	HB	HG12	HG13
ILE 1	1	2	3												

Abbildung 30: Backbone(N,Cα,C′)-Template. Die korrespondierenden Atome N, Cα und C′ sind auf 3 Distanzmatrizen verteilt. Die Zahlen in den grünen Zellen entsprechen den Nummern der Distanzmatrizen. Somit bildet z.B. die erste Distanzmatrix die Korrespondenzen zwischen den Stickstoffen ab.

jeweiligen N-, Cα- und C′-Atomen. Die Unspecific-Substitutionsmatrix (Abb. 31) definiert die Substitution eines Residuums durch ein beliebiges, als gleich wahrscheinlich. Im ersten Vergleichsmodus Backbone(N,Cα,C′) / Unspecific ermittelt die neue Version (Tab. 8) neben einer CS mit 138 Residuentupeln ($RMSD = 1.419\text{Å}$), die auch von der alten Version erkannt worden ist, drei MCSs mit je 139 Residuentupeln. Die MCS ist in diesem Fall auch die BCS ($NBCSS = 1.0$). Während die alte Version für diese Aufgabe 173*s* auf einem Kern des Intel Core 2 Quad Q6600 gebraucht hat, beträgt die Laufzeit der neuen Version 2.7*s* auf einem Kern des Xeon X5650, wobei die Rechenzeit je nach Einstellung des Threshold-Korridors variieren kann. Ausgehend davon, dass die Leistung eines Kerns des Xeon X5650 etwa doppelt so hoch ist als die Leistung eines Kerns des Intel Core 2 Quad Q6600, zeigt

substitution matrix

20 / 20 / 400 / 400

20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	ALA 1	ARG 1	ASN 1	ASP 1	CYS 1	GLN 1	GLU 1	GLY 1	HIS 1	ILE 1	LEU 1	LYS 1	MET 1	PHE 1	PRO 1	SER 1	THR 1	TRP 1	TYR 1	VAL 1
ALA 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ARG 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ASN 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ASP 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CYS 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
GLN 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
GLU 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
GLY 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HIS 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ILE 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
LEU 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
LYS 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MET 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
PHE 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
PRO 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SER 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
THR 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
TRP 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
TYR 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
VAL 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Abbildung 31: UNSPECIFIC-Substitutionsmatrix. Die Substitution eines Residuums durch ein beliebiges Residuum ist gleich wahrscheinlich und somit absolut unspezifisch.

COMPS	ATOMS	RMSD	NWRMSD	IDENT	SSIM	CSS	QMCSS	TMCSS	QTMCSS	NBCSS	MCSS
139	417	1.44	0.834	0.439	1.0	0.914	0.901	0.87	0.885	1.0	0.901
139	417	1.515	0.82	0.439	1.0	0.914	0.901	0.87	0.885	0.983	0.901
139	417	1.583	0.807	0.439	1.0	0.913	0.9	0.87	0.885	0.968	0.9
138	414	1.419	0.838	0.442	1.0	0.915	0.896	0.865	0.88	0.997	0.896
138	414	1.425	0.837	0.442	1.0	0.915	0.896	0.865	0.88	0.996	0.896
138	414	1.496	0.823	0.442	1.0	0.915	0.895	0.865	0.88	0.981	0.895

Tabelle 8: Deskriptoren der BCS und der MCSs im unspezifischen Fall. Je drei Alternativalignments pro Ordnungen $k = 139$ und $k = 138$.

dieses Ergebnis eine rund 32-fache Steigerung der Performance in Verbindung mit einer Verbesserung der Qualität. Es sei an dieser Stelle darauf hingewiesen, dass die Unterschiede in den Ähnlichkeitsangaben zu den gleichen Strukturen (Tab. 4 mit $MCSS = 0.845$ und Tab. 8 mit $MCSS = 0.901$) sich aus unterschiedlichen Residuen-Templates und Substitutionsmatrizen ergeben. So führt die Berücksichtigung der Atome der Residuenreste, und somit der konformationellen Änderungen der Rotamere, zu einer größeren RMSD. Zusätzlich erlaubt die Verwendung einer BLOSUM62-Substitutionsmatrix die Bewertung der Substitutionsähnlichkeit (Tab. 4). In diesem Beispiel (Tab. 8) wird die Substitutionsähnlichkeit gänzlich ignoriert und die alleinige Berücksichtigung von lediglich 3 Rückgratatomen pro Residuum bewertet die Ähnlichkeit der Rückgratgeometrie nur unvollständig. Daher verfügt der Vergleichsmodus ALLATOMS / BLOSUM62-SIGMOID (Abb. 5 / Abb. 7) über einen deutlich höheren Informationsgehalt und bewertet die Ähnlichkeit der beiden Strukturen realistischer. Das zweite Beispiel in der Publikation ist der Vergleich der beiden Ketten im Modus BACKBONE(Cα,Cβ) / SPECIFIC. Das BACKBONE(Cα,Cβ)-Template (Abb. 32) berücksichtigt die interatomaren Distanzen zwischen den jeweiligen Cα- und Cβ-Atomen. Die SPECIFIC-Substitutionsmatrix (Abb. 33) erlaubt die Substitution eines Residuums nur durch eines des gleichen Typs. EPITOPEMATCH markiert die Permutation mit 61 Residuen als BCS (Tab. 9). Die neue Version terminiert nach 1.18*s* (Xeon X5650) und erreicht eine 22-fache Performancesteigerung gegenüber der Laufzeit von 52.46*s* (Intel Core 2 Quad Q6600) der alten Version. Abb. 34a zeigt das ermittelte Ähnlichkeitsspektrum. Die Identität und die Substitutionsähnlichkeit (IDENT, SSIM, blau) sind im Modus SPECIFIC durchwegs gleich 1.0. Die Query-, Target-, und Query/Target-Ähnlichkeit (QMCSS, TMCSS, QTMCSS, magenta, pink)

components

20 / 20 / 14810

ALA
ARG
ASN
ASP
CYS
GLN
GLU
GLY
HIS
ILE
LEU
LYS
MET
PHE
PRO
SER
THR
TRP
TYR
VAL

components and their atoms

0 / 20 0 / 387 20 / 20 39 / 387

20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ALA 1	N	CA	C	O	CB	OXT	H	H2	HA	HB1	HB2	HB3	HXT		
ALA 1		1			2										
ARG 1	N	CA	C	O	CB	CG	CD	NE	CZ	NH1	NH2	OXT	H	H2	HA
ARG 1		1			2										
ASN 1	N	CA	C	O	CB	CG	OD1	ND2	OXT	H	H2	HA	HB2	HB3	HD21
ASN 1		1			2										
ASP 1	N	CA	C	O	CB	CG	OD1	OD2	OXT	H	H2	HA	HB2	HB3	HD2
ASP 1		1			2										
CYS 1	N	CA	C	O	CB	SG	OXT	H	H2	HA	HB2	HB3	HG	HXT	
CYS 1		1			2										
GLN 1	N	CA	C	O	CB	CG	CD	OE1	NE2	OXT	H	H2	HA	HB2	HB3
GLN 1		1			2										
GLU 1	N	CA	C	O	CB	CG	CD	OE1	OE2	OXT	H	H2	HA	HB2	HB3
GLU 1		1			2										
GLY 1	N	CA	C	O	OXT	H	H2	HA2	HA3	HXT					
GLY 1		1													
HIS 1	N	CA	C	O	CB	CG	ND1	CD2	CE1	NE2	OXT	H	H2	HA	HB2
HIS 1		1			2										
ILE 1	N	CA	C	O	CB	CG1	CG2	CD1	OXT	H	H2	HA	HB	HG12	HG13
ILE 1		1			2										

Abbildung 32: BACKBONE(Cα,Cβ)-Template. Die korrespondierenden Atome Cα und Cβ sind auf 2 Distanzmatrizen verteilt. Die Zahlen in den grünen Zellen entsprechen den Nummern der Distanzmatrizen. Für GLY existieren keine Korrespondenzen in der zweiten Distanzmatrix. Die Substitution von GLY durch ein nicht GLY wird in diesem Fall durch ein Penalty der Komponentenvollständigkeit $COMPL(RT^{QT}_{G,\mathcal{G}}) = \frac{1}{2} = 0.5$ bestraft.

substitution matrix

20 / 20 / 20 / 400

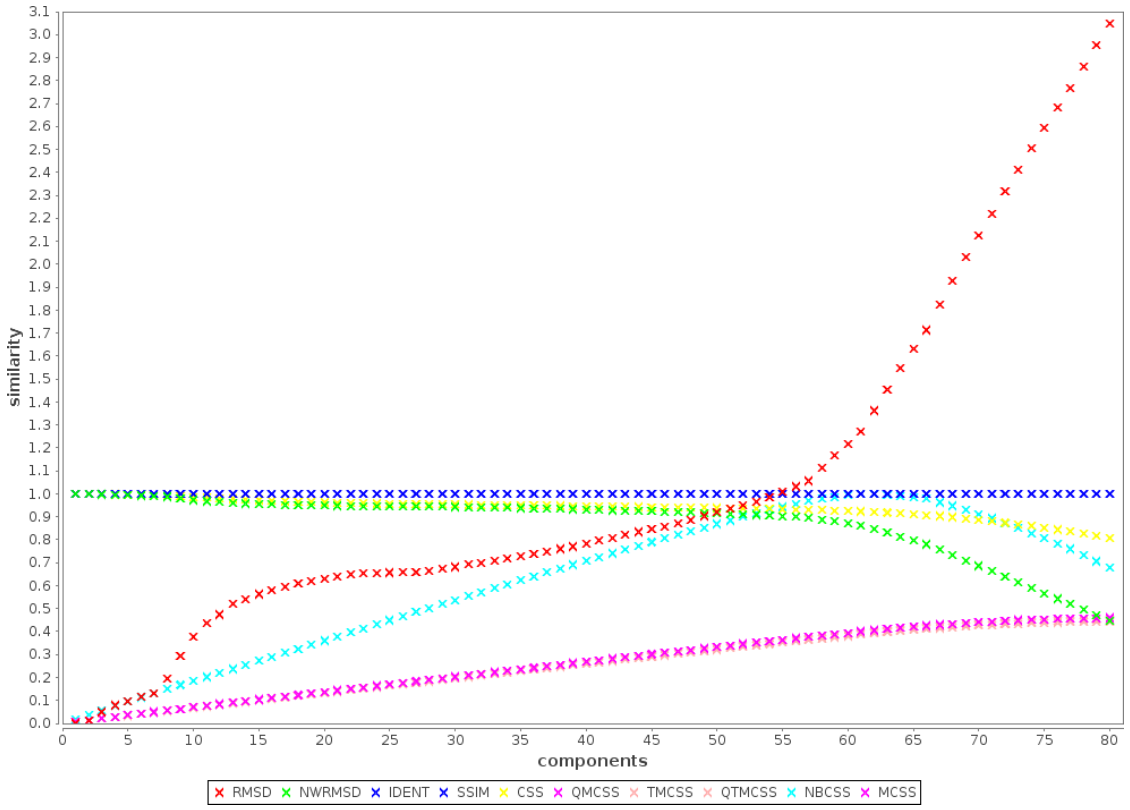
20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	ALA 1	ARG 1	ASN 1	ASP 1	CYS 1	GLN 1	GLU 1	GLY 1	HIS 1	ILE 1	LEU 1	LYS 1	MET 1	PHE 1	PRO 1	SER 1	THR 1	TRP 1	TYR 1	VAL 1
ALA 1	1.000																			
ARG 1		1.000																		
ASN 1			1.000																	
ASP 1				1.000																
CYS 1					1.000															
GLN 1						1.000														
GLU 1							1.000													
GLY 1								1.000												
HIS 1									1.000											
ILE 1										1.000										
LEU 1											1.000									
LYS 1												1.000								
MET 1													1.000							
PHE 1														1.000						
PRO 1															1.000					
SER 1																1.000				
THR 1																	1.000			
TRP 1																		1.000		
TYR 1																			1.000	
VAL 1																				1.000

Abbildung 33: SPECIFIC-Substitutionsmatrix. Die Substitution eines Residuums durch ein Residuum eines anderen Typs ist unzulässig und somit absolut spezifisch.

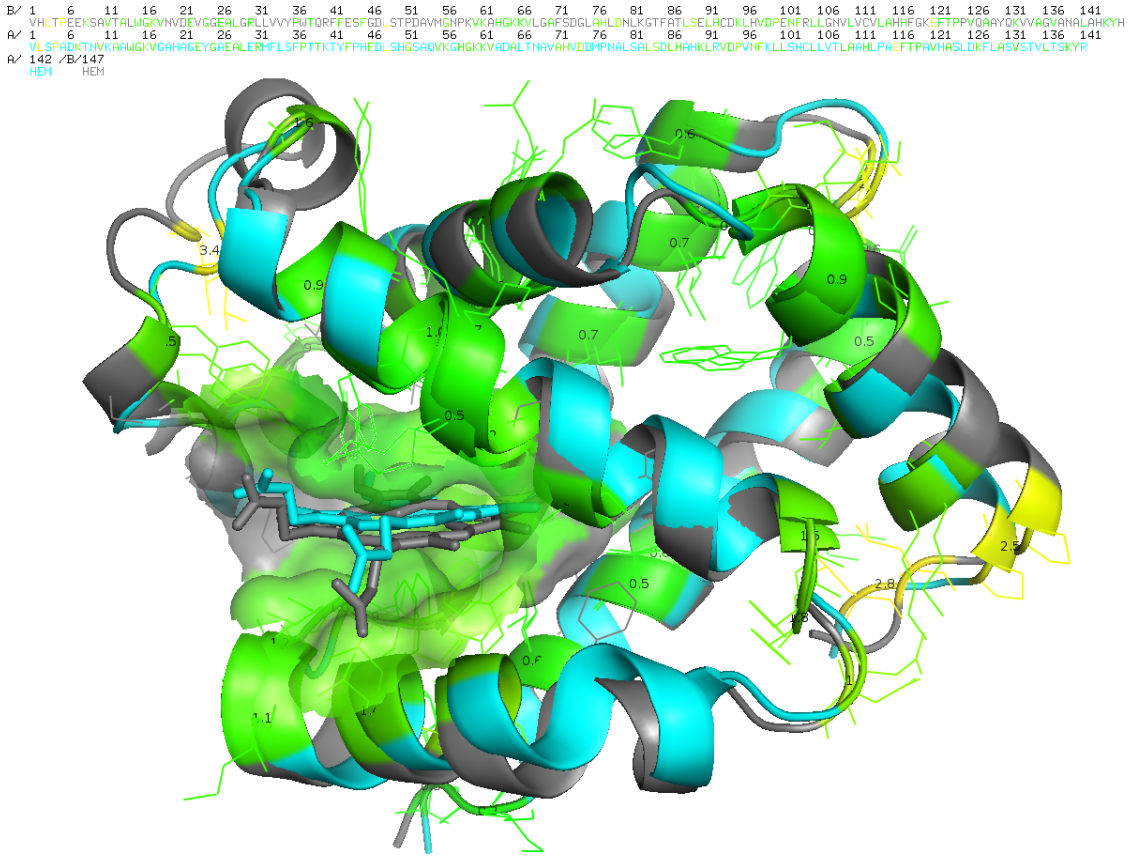
COMPS	ATOMS	RMSD	NWRMSD	IDENT	SSIM	CSS	QMCSS	TMCSS	QTMCSS	NBCSS	MCSS
80	155	3.047	0.446	1.0	1.0	0.809	0.459	0.443	0.451	0.678	0.459
79	153	2.954	0.471	1.0	1.0	0.818	0.458	0.443	0.45	0.706	0.458
...											
62	120	1.362	0.848	1.0	1.0	0.921	0.405	0.391	0.398	0.998	0.405
61	118	1.27	0.863	1.0	1.0	0.923	0.399	0.386	0.392	1.0	0.399
60	116	1.218	0.872	1.0	1.0	0.925	0.394	0.38	0.387	0.993	0.394

Tabelle 9: Deskriptoren der BCS und der MCS im spezifischen Fall. Die BCS ist 61 und die MCS ist 62 Residuen groß. Die CSs oberhalb dieser Größen sind falsch-positiv.

sind relativ gleich, da die beiden Ketten (A, 141 und B, 146 Residuen) ungefähr gleich groß sind. QMCSS ist in diesem Fall gleich MCSS (magenta), weil die QS (Kette A) kleiner ist. Die ab



(a) Deskriptoren im Modus BACKBONE(Cα,Cβ) / SPECIFIC. Der Scheitelpunkt der NBCSS-Parabel ($k = 61$, cyan) markiert die BCS. Der steile Anstieg der RMSD ab der Ordnung $k = 63$ zeugt von einer falsch-positiven Entwicklung.



(b) Die BCS der 2DN2.A (cyan) und der 2DN2.B (grau) im Modus BACKBONE(Cα,Cβ) / SPECIFIC. Die korrespondierenden Residuen sind nach der $NWRMSD(RT_{qt}^{QT})$ (min max) gefärbt. Die höchste Konzentration der identischen Residuen ist im Bereich der Liganden.

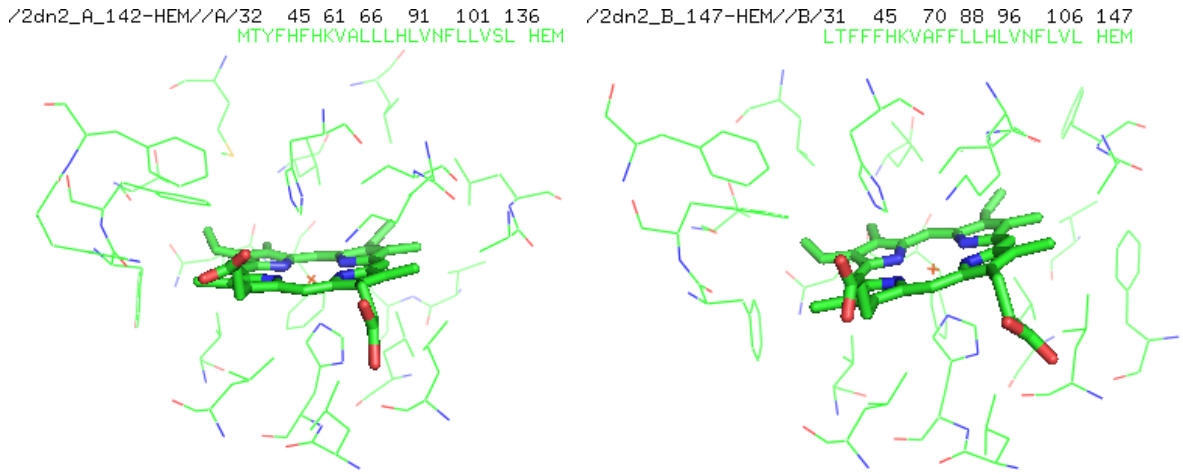
Abbildung 34: Modus BACKBONE(Cα,Cβ) / SPECIFIC.

dem Residuentupel 62 steil ansteigende RMSD resultiert aus der Zuordnung von Residuentupeln, die eher dem geometrischen Rauschen zuzuschreiben sind, sodass diese geometrische Abweichung im Bezug auf die Größe der QS (37) den Verlauf der BCS-Parabel (cyan) am Residuentupel 62 umkehren lässt. Die BCS markiert in diesem Vergleichsmodus den spezifischen Ähnlichkeitskern der beiden Ketten. Abb. 34b zeigt die BCS (Tab. 9, $COMPS = 61$). Es ist gut zu erkennen, dass die Konzentration der identischen Residuen der beiden Ketten in der Region der Liganden am höchsten ist. Somit kann man annehmen, dass die physikochemische Ähnlichkeit der beiden Strukturen in der Region der Epitope am höchsten konserviert ist. Während die spezifische (Abb. 33) Homologie der beiden Ketten bei 39.9% für die BCS und 45.9% für die MCS liegt (Tab. 9), erreicht die Bewertung der gewichteten (Abb. 7) Homologie deutlich höheren Werte, 82.7% für die BCS und 84.5% für die MCS (Tab. 4). Die spezifische Homologie ist eine Teilmenge der gewichteten Homologie. Der Anteil der identischen Aminosäuren im gewichteten Vergleich (Tab. 4, $IDENT = 0.439$) entspricht mit $round(0.439 \cdot 141) = 62$ Aminosäuren der Größe der zur BCS unmittelbar benachbarten CS (Tab. 9). Die Differenz der NBCSSs der beiden gemeinsamen Substrukturen beträgt lediglich $1.0 - 0.998 = 0.002$. Alle spezifisch gematchten Residuentupel der Ordnungen ≥ 63 sind im gewichteten Match nicht enthalten, und können somit als falsch-positive Zuordnungen betrachtet werden.

2.3.1.2 *Epitope der homologen Strukturen*

2DN2

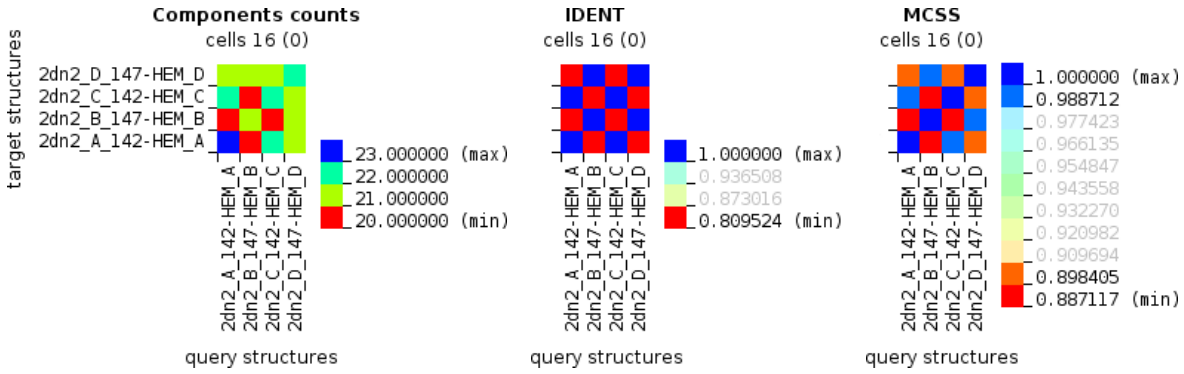
Jede der 4 Ketten aus 2DN2 ist ein Protein-Ligand-Komplex (Apo-Struktur) und bindet ein Protoporphyrin IX mit Fe (HEM) [129]. EpitopeMatch ist mit dem Auschneidewerkzeug ausgestattet, mit dem die Umgebungen von definierten Residuen bzw. Liganden aus den Strukturen ausgeschnitten werden können. Die Abb. 35 zeigt die ausgeschnittenen Residuen



(a) HEM-5$\AA$-Umgebung der Kette A. (b) HEM-5$\AA$-Umgebung der Kette B.

Abbildung 35: HEM-5$\AA$-Umgebungen der Ketten A und B aus 2DN2

der Kette A und B, die mit mindestens einem Atom in die 5$\AA$-Umgebung des jeweiligen HEM hineinragen. Bei den Distanzmessungen zwischen den Atomen der Aminosäuren und des Liganden sind ausschließlich schwere Atome (keine Wasserstoffatome) berücksichtigt worden. Die Abb. 36 zeigt den Kreuzvergleich der Epitope aller 4 Ketten. Die Anzahl der Residuen pro Epitop kann von der Diagonale der Matrix mit der Anzahl der gematchten Residuen (Abb. 36, a) entnommen werden. Demnach befinden sich 23 Aminosäuren der



(a) Anzahl der Residuen. (b) Identität. (c) Ähnlichkeit.

Abbildung 36: Kreuzvergleich der Epitope aus 2DN2.

Kette A, 21 Aminosäuren der Kette B und jeweils 22 Aminosäuren der Ketten C und D in den 5$\mathring{A}$-Umgebungen der jeweiligen HEM-Moleküle. Die dargestellten Kreuzvergleich-Matrizen sind symmetrisch. Das Match der Epitope der Ketten A und C besteht aus 22 Residuentupel. Das Match der Epitope der Ketten B und D besteht aus 21 Residuentupel. Der Deskriptor IDENT (Abb. 36, b) zeigt, dass alle 22 Residuentupel die Epitope der Ketten A und C aus identischen Aminosäuren bestehen. Das gleiche gilt für alle 21 Residuentupel der Epitope der Ketten B und D. Dies ist ein Indikator dafür, dass die Ketten A und C bzw. die Ketten B und D identisch sind. Obwohl es in diesem Fall auch so ist, bedeutet die hundertprozentige Identität der Epitope nicht zwangsläufig die hundertprozentige Identität der Strukturen, die sie tragen. Die Identität der homologen Epitope der Ketten A und B bzw. der Ketten C und D liegt bei $\approx 80,95\%$. Dies bestätigt die Annahme, dass die physikochemische Ähnlichkeit in den Regionen der Epitope höher konserviert ist als dieselbe der vollständigen Ketten (Abs. 2.2.5). Mit $\approx 80,95\%$ ist die Identität der homologen Epitope fast zwei Mal höher als die Identität der homologen Strukturen mit $\approx 43.9\%$. Die Abb. 36, c zeigt die Verteilung der Ähnlichkeit der Epitope. Die Bindungsstellen ähneln sich mindestens zu $\approx 88,71\%$. Der Kreuzvergleich nahm $1.77s$ in Anspruch.

Die Tab. 10 zeigt die Ergebnisse der Suche nach allen Epitopen auf allen Ketten. Jedes Epi-

QUERY	TARGET	COMPS	ATOMS	RMSD	NWRMSD	IDENT	SSIM	CSS	QMCSS	TMCSS	QTMCSS	NBCSS	MCSS
2dn2_A_142-HEM_A	2dn2_A	23	169	0.0	1.0	1.0	0.989	1.0	1.0	0.163	0.28	1.0	1.0
	2dn2_B	23	164	1.088	0.885	0.783	0.917	0.922	0.922	0.145	0.251	1.0	0.922
	2dn2_C	23	169	0.361	0.974	1.0	0.989	0.99	0.99	0.162	0.278	1.0	0.99
	2dn2_D	23	164	0.995	0.898	0.783	0.917	0.931	0.931	0.147	0.253	1.0	0.931
2dn2_B_147-HEM_B	2dn2_A	21	153	1.065	0.887	0.81	0.934	0.922	0.922	0.137	0.239	1.0	0.922
	2dn2_B	21	157	0.0	1.0	1.0	0.99	1.0	1.0	0.144	0.251	1.0	1.0
	2dn2_C	21	153	1.003	0.896	0.81	0.934	0.93	0.93	0.138	0.241	1.0	0.93
	2dn2_D	21	157	0.352	0.975	1.0	0.99	0.989	0.989	0.142	0.249	1.0	0.989
2dn2_C_142-HEM_C	2dn2_A	22	163	0.366	0.973	1.0	0.989	0.99	0.99	0.154	0.267	1.0	0.99
	2dn2_B	22	159	0.983	0.902	0.818	0.926	0.931	0.931	0.14	0.244	1.0	0.931
	2dn2_C	22	163	0.0	1.0	1.0	0.989	1.0	1.0	0.156	0.27	1.0	1.0
	2dn2_D	22	159	0.892	0.914	0.818	0.926	0.94	0.94	0.142	0.246	1.0	0.94
2dn2_D_147-HEM_D	2dn2_A	22	159	0.971	0.899	0.773	0.904	0.931	0.931	0.145	0.251	1.0	0.931
	2dn2_B	22	163	0.349	0.975	1.0	0.989	0.989	0.989	0.149	0.259	1.0	0.989
	2dn2_C	22	159	0.912	0.908	0.773	0.904	0.939	0.939	0.147	0.254	1.0	0.939
	2dn2_D	22	163	0.0	1.0	1.0	0.989	1.0	1.0	0.151	0.262	1.0	1.0

Tabelle 10: Epitope aus 2DN2 vs. Ketten aus 2DN2.

top ist auf jeder identischen und jeder homologen Kette vollständig erkannt. Der Run wurde im Modus ALLATOMS / BLOSUM62-SIGMOID durchgeführt. Die Rechenzeit betrug $3.771s$. Die Ähnlichkeit der Epitope der homologen Ketten bewegt sich zwischen $MCSS \approx 92.2\%$ (A, B) und $MCSS \approx 94.0\%$ (C, D). Die Ähnlichkeit der Epitope der identischen Ketten

bewegt sich zwischen $MCSS \approx 98.9\%$ (B, D) und $MCSS \approx 99.0\%$ (A, C). Die Ähnlichkeit der homologen Ketten ist niedriger, da zu der Konformationsänderung die geringere physiko-chemische Ähnlichkeit durch Residuenpaare mit nicht identischen Aminosäuren hinzukommt.

1Q1A & 1MA3

Im zweiten Anwendungsbeispiel wird das Epitop eines Hst2-Hefe-Proteins (Struktur 1Q1A [188]) mit dem Epitop eines bakteriellen Sir2-Proteins (Struktur 1MA3 [16]) verglichen. Beide Proteine binden das Co-Substrat ADP-Ribose und übertragen eine Acetylgruppe von einem Peptid-Substrat auf die ADP-Ribose, wobei die 2'-O-Acetyl-ADP-Ribose (OAD) entsteht. 1Q1A liegt im Komplex mit OAD vor. 1MA3 ist hingegen eine Apo-Struktur. Während in der Publikation zu EpitopeMatch [77] das Augenmerk allein auf der Erkennung des OAD-Epitops gelegt wird, konzentriert sich die neue Version auf die Analyse der Homologie der beiden Strukturen und ihrer Epitope. Als Neuerung gegenüber der ersten Version stellt EpitopeMatch neben dem AllAtoms-Template (Abb. 5), das zum deutlich genaueren Erfassen der geometrische Ähnlichkeit führt, eine Substitutionsmatrix (Abb. 7) zur Verfügung, mit deren Hilfe die physiko-chemische Ähnlichkeit charakterisiert werden kann. Die Substitutionsmatrizen [42, 61] geben pro Aminosäurenpaar eine Substitutionswahrscheinlichkeit an und unterscheiden sich in der Betrachtung der relativen Mutationsraten verwandter Proteine. Die Substitutionswahrscheinlichkeiten implizieren zwar die Information über die physiko-chemische Ähnlichkeit der Aminosäuren, allerdings können sie durch weitere physiko-chemische Eigenschaften wie AVErage HYDROPathy (AVEHYDROP) [99]

amino acid	MOLWEIGHT	AVEHYDROP	NETCHARGE(pH 7)	$pK_1(COOH)$	$pK_2(NH_3^+)$	pK_R(R group)
ALA (A)	71.0788	1.8	-0.002016	2.34	9.69	
ARG (R)	156.1875	-4.5	0.990974	2.17	9.04	12.48
ASN (N)	114.1038	-3.5	-0.015591	2.02	8.80	
ASP (D)	115.0886	-3.5	-1.002052	1.88	9.60	3.65
CYS (C)	103.1388	2.5	-0.062490	1.96	10.28	8.18
GLN (Q)	128.1307	-3.5	-0.007344	2.17	9.13	
GLU (E)	129.1155	-3.5	-1.000343	2.19	9.67	4.25
GLY (G)	57.0519	-0.4	-0.002484	2.34	9.60	
HIS (H)	137.1411	-3.2	0.084200	1.82	9.17	6.00
ILE (I)	113.1594	4.5	-0.002062	2.36	9.68	
LEU (L)	113.1594	3.8	-0.002483	2.36	9.60	
LYS (K)	128.1741	-3.9	0.988624	2.18	8.95	10.53
MET (M)	131.1926	1.9	-0.006109	2.28	9.21	
PHE (F)	147.1766	2.8	-0.007352	1.83	9.13	
PRO (P)	97.1167	-1.6	-0.000100	1.99	10.96	
SER (S)	87.0782	-0.8	-0.007013	2.21	9.15	
THR (T)	101.1051	-0.7	-0.002380	2.11	9.62	
TRP (W)	186.2132	-0.9	-0.004033	2.38	9.39	
TYR (Y)	163.1760	-1.3	-0.008537	2.20	9.11	10.07
VAL (V)	99.1326	4.2	-0.002372	2.32	9.62	

Tabelle 11: Physiko-chemischen Eigenschaften der Aminosäuren.

und MOLecular WEIGHT (MOLWEIGHT) [180, 50] ergänzt werden (Tab. 11). Die net charge (NETCHARGE) wird nach [75]

$$Z = \sum_i N_i \frac{10^{pKa_i}}{10^{pH} + 10^{pKa_i}} - \sum_j N_j \frac{10^{pH}}{10^{pH} + 10^{pKa_j}}$$

mit

- i Index für N-Terminus und ARG, LYS, HIS.
- j Index für C-Terminus und ASP, GLU, CYS, TYR.

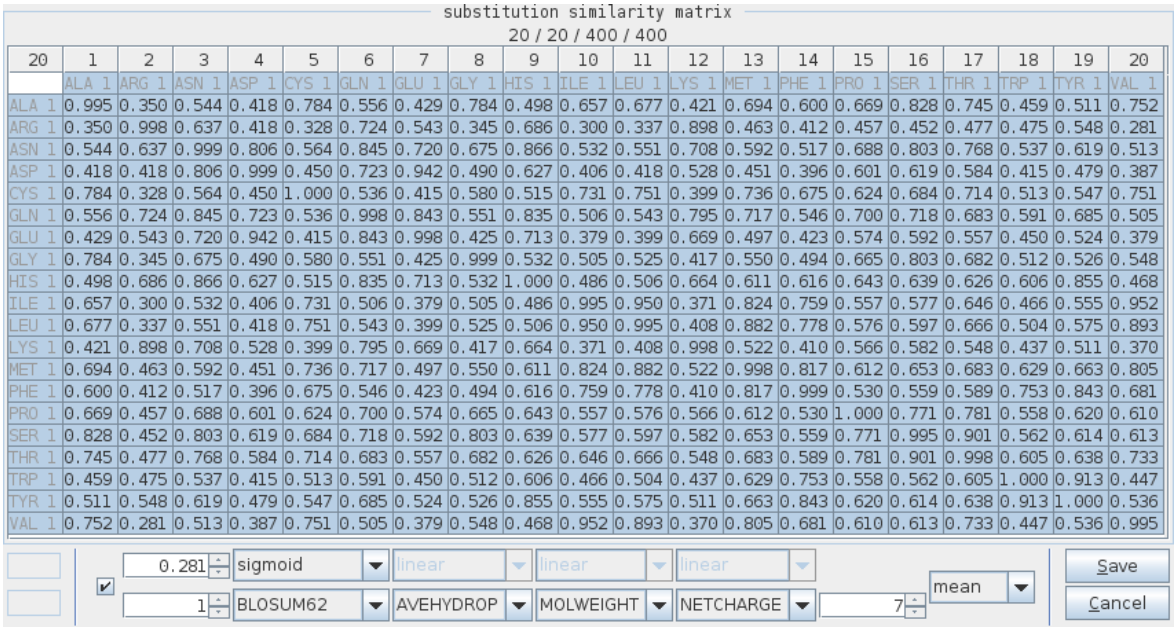
und den pK_a-Werten aus [122] berechnet. Beispiele für pH 7.0:

$$Z(ALA) = \frac{10^{9.69}}{10^{7.0}+10^{9.69}} - \frac{10^{7.0}}{10^{7.0}+10^{2.34}} \approx -0.002016$$

$$Z(ARG) = \frac{10^{9.04}}{10^{7.0}+10^{9.04}} + \frac{10^{12.48}}{10^{7.0}+10^{12.48}} - \frac{10^{7.0}}{10^{7.0}+10^{2.17}} \approx 0.990974$$

$$Z(ASP) = \frac{10^{9.6}}{10^{7.0}+10^{9.6}} - \frac{10^{7.0}}{10^{7.0}+10^{1.88}} - \frac{10^{7.0}}{10^{7.0}+10^{3.65}} = -1.002052$$

Die Abb. 37 zeigt die Substitutionsmatrix, in der die sigmoid normalisierten Substitutions-



20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	ALA 1	ARG 1	ASN 1	ASP 1	CYS 1	GLN 1	GLU 1	GLY 1	HIS 1	ILE 1	LEU 1	LYS 1	MET 1	PHE 1	PRO 1	SER 1	THR 1	TRP 1	TYR 1	VAL 1
ALA 1	0.995	0.350	0.544	0.418	0.784	0.556	0.429	0.784	0.498	0.657	0.677	0.421	0.694	0.600	0.669	0.828	0.745	0.459	0.511	0.752
ARG 1	0.350	0.998	0.637	0.418	0.328	0.724	0.543	0.345	0.686	0.300	0.337	0.898	0.463	0.412	0.457	0.452	0.477	0.475	0.548	0.281
ASN 1	0.544	0.637	0.999	0.806	0.564	0.845	0.720	0.675	0.866	0.532	0.551	0.708	0.592	0.517	0.688	0.803	0.768	0.537	0.619	0.513
ASP 1	0.418	0.418	0.806	0.999	0.450	0.723	0.942	0.490	0.627	0.406	0.418	0.528	0.451	0.396	0.601	0.619	0.584	0.415	0.479	0.387
CYS 1	0.784	0.328	0.564	0.450	1.000	0.536	0.415	0.580	0.515	0.731	0.751	0.399	0.736	0.675	0.624	0.684	0.714	0.513	0.547	0.751
GLN 1	0.556	0.724	0.845	0.723	0.536	0.998	0.843	0.551	0.835	0.506	0.543	0.795	0.717	0.546	0.700	0.718	0.683	0.591	0.685	0.505
GLU 1	0.429	0.543	0.720	0.942	0.415	0.843	0.998	0.425	0.713	0.379	0.399	0.669	0.497	0.423	0.574	0.592	0.557	0.450	0.524	0.379
GLY 1	0.784	0.345	0.675	0.490	0.580	0.551	0.425	0.999	0.532	0.505	0.525	0.417	0.550	0.494	0.665	0.803	0.682	0.512	0.526	0.548
HIS 1	0.498	0.686	0.866	0.627	0.515	0.835	0.713	0.532	1.000	0.486	0.506	0.664	0.611	0.616	0.643	0.639	0.626	0.606	0.855	0.468
ILE 1	0.657	0.300	0.532	0.406	0.731	0.506	0.379	0.505	0.486	0.995	0.950	0.371	0.824	0.759	0.557	0.577	0.646	0.466	0.555	0.952
LEU 1	0.677	0.337	0.551	0.418	0.751	0.543	0.399	0.525	0.506	0.950	0.995	0.408	0.882	0.778	0.576	0.597	0.666	0.504	0.575	0.893
LYS 1	0.421	0.898	0.708	0.528	0.399	0.795	0.669	0.417	0.664	0.371	0.408	0.998	0.522	0.410	0.566	0.582	0.548	0.437	0.511	0.370
MET 1	0.694	0.463	0.592	0.451	0.736	0.717	0.497	0.550	0.611	0.824	0.882	0.522	0.998	0.817	0.612	0.653	0.683	0.629	0.663	0.805
PHE 1	0.600	0.412	0.517	0.396	0.675	0.546	0.423	0.494	0.616	0.759	0.778	0.410	0.817	0.999	0.530	0.559	0.589	0.753	0.843	0.681
PRO 1	0.669	0.457	0.688	0.601	0.624	0.700	0.574	0.665	0.643	0.557	0.576	0.566	0.612	0.530	1.000	0.771	0.781	0.558	0.620	0.610
SER 1	0.828	0.452	0.803	0.619	0.684	0.718	0.592	0.803	0.639	0.577	0.597	0.582	0.653	0.559	0.771	0.995	0.901	0.562	0.614	0.613
THR 1	0.745	0.477	0.768	0.584	0.714	0.683	0.557	0.682	0.626	0.646	0.666	0.548	0.683	0.589	0.781	0.901	0.998	0.605	0.638	0.733
TRP 1	0.459	0.475	0.537	0.415	0.513	0.591	0.450	0.512	0.606	0.466	0.504	0.437	0.629	0.753	0.558	0.562	0.605	1.000	0.913	0.447
TYR 1	0.511	0.548	0.619	0.479	0.547	0.685	0.524	0.526	0.855	0.555	0.575	0.511	0.663	0.843	0.620	0.614	0.638	0.913	1.000	0.536
VAL 1	0.752	0.281	0.513	0.387	0.751	0.505	0.379	0.548	0.468	0.952	0.893	0.370	0.805	0.681	0.610	0.613	0.733	0.447	0.536	0.995

Abbildung 37: Substitutionsmatrix B62sAhMwNc7. Die physiko-chemischen Eigenschaften aus Tab. 11 werden in Form von normalisierten Differenzen gemittelt (mean). Der Benutzer kann sie wahlweise aktivieren/deaktivieren, um andere Kombinationen einzustellen. Neben dem Mittelwert kann auch ein Produkt ausgewählt werden. Der Produkt der normalisierten Differenzen führt zu einer spezifischeren Substitutionsmatrix, in der die Kombinationen mit den maximalen Differenzen ARG/LEU (AVEHYDROP), GLY/TRP (MOLWEIGHT) und ARG/ASP (NETCHARGE) mit 0.0 gewichtet werden. Die Gewichte aus der Substitutionsmatrix fließen unmittelbar in die Berechnung der WRMSD (Gl. 33) bzw. der SSIM (Gl. 44).

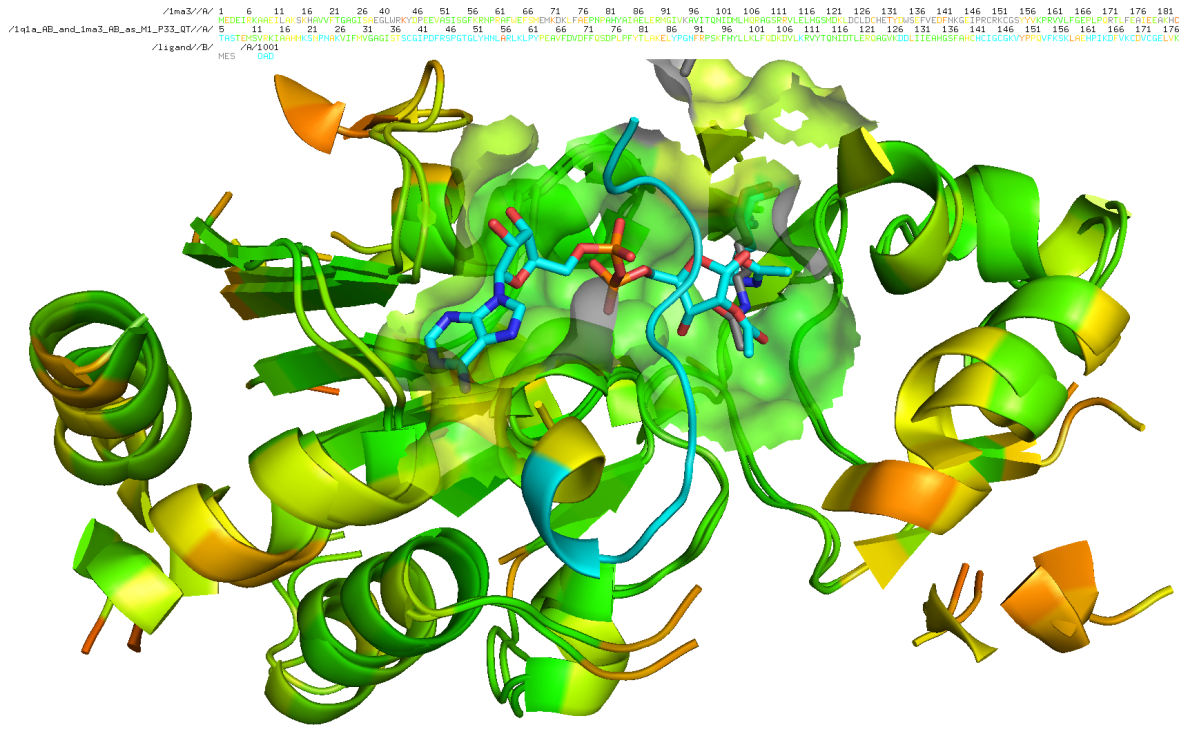
wahrscheinlichkeiten der Blosum62-Matrix mit der AVEHYDROP, dem MOLWEIGHT und der NETCHARGE beim PH-Wert 7 als Mittelwert der normalisierten Differenzen kombiniert sind:

$$ss_{ij}^{B62S,AH,MW,NC7} = \frac{ss_{ij}^{BM_{sigmoid}^{62}} + 3 - \frac{|\text{diff}_{ij}^{AH}|}{\text{diff}_{max}^{AH}} - \frac{|\text{diff}_{ij}^{MW}|}{\text{diff}_{max}^{MW}} - \frac{|\text{diff}_{ij}^{NC7}|}{\text{diff}_{max}^{NC7}}}{4} \tag{63}$$

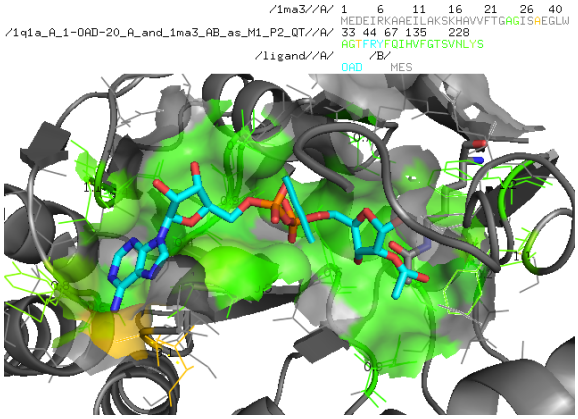
Als ungünstigste Substitution stellt sich in diesem Fall das Aminosäurenpaar ARG,VAL

$$ss_{ARG,VAL}^{B62S,AH,MW,NC7} = \frac{0.03 + 3 - \frac{|-4.5-4.2|}{|-4.5-4.5|} - \frac{|99.1326-156.1875|}{|57.0519-186.2132|} - \frac{|-0.002372-0.990974|}{|-1.002052-0.990974|}}{4} \approx 0.281$$

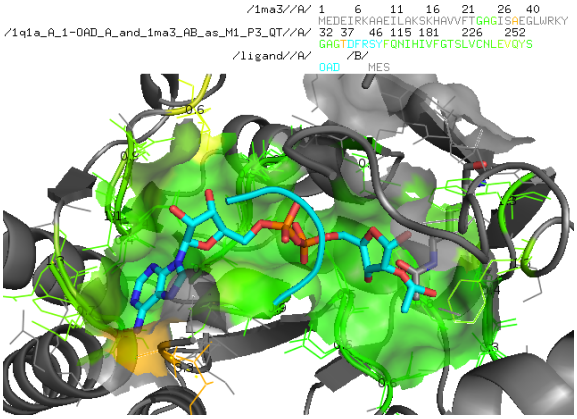
heraus, mit lediglich $\approx$ 28.1%, gefolgt von ARG,ILE mit $\approx$ 30.0% und ARG,CYS mit $\approx$ 32.8%. Die Template/Substitutionsmatrix-Kombination AllAtoms/B62sAhMwNc7 ist die Standardeinstellung von EpitopeMatch. Abb. 38a zeigt die beste gemeinsame Substruk-



(a) BCS der Strukturen 1Q1A & 1MA3.



(b) 20 Aminosäuren des Epitops.



(c) 5Å-Umgebung mit 30 Aminosäuren.

Abbildung 38: BCS und Epitop der 1Q1A & 1MA3. Die Färbung erfolgt nach der $NWRMSD(RT_{qt}^{QT})$ (min max) der korrespondierenden Residuen.

tur ($NBCSS = 1.0$) der beiden Deacetylasen. Die beiden Strukturen (Query/1Q1A/cyan; Target/1MA3/grau) verfügen über ein gut konserviertes gemeinsames Faltungsmuster (grün). Die Tertiärstruktur des Faltungsmusters besteht aus sechs parallel verlaufenden β-Faltblättern (Mitte/links), die auf der einer Seite von vier α-Helices (vorne/links) und auf der anderen Seite von zwei α-Helices (hinten/links) umgeben sind. Das OAD wird von den Aminosäuren der Random-Coils gebunden, die diese Sekundärstrukturen verbinden. Das vierte β-Faltblatt (von links) mündet in einem kurzen helikalen Abschnitt, der in einen Loop übergeht. Der Loop (Mitte/oben/cyan) hat die Funktion den Liganden in der Bindetasche zu fixieren oder ihn wieder frei zu lassen. Der Loop der Apo-Struktur konnte

aufgrund seiner Flexibilität nicht kristallisiert werden. Die acetylierten Lysin-Seitenketten, die mit den entsprechenden Peptiden (Histon H4 und zelluläres Tumorantigen p53) sowohl mit der Holo-, als auch mit der Apo-Struktur kristallisiert worden sind, liegen mit ihren Acetyl-Gruppen in der unmittelbaren Nähe der Acetyl-Gruppe des OAD. Dies betont die Richtigkeit des Match und den Fakt, dass beide Strukturen über ein gemeinsames, starres Faltungsmuster verfügen.

Abb. 39 zeigt den Verlauf der Deskriptoren der gemeinsamen Substrukturen bzw. den

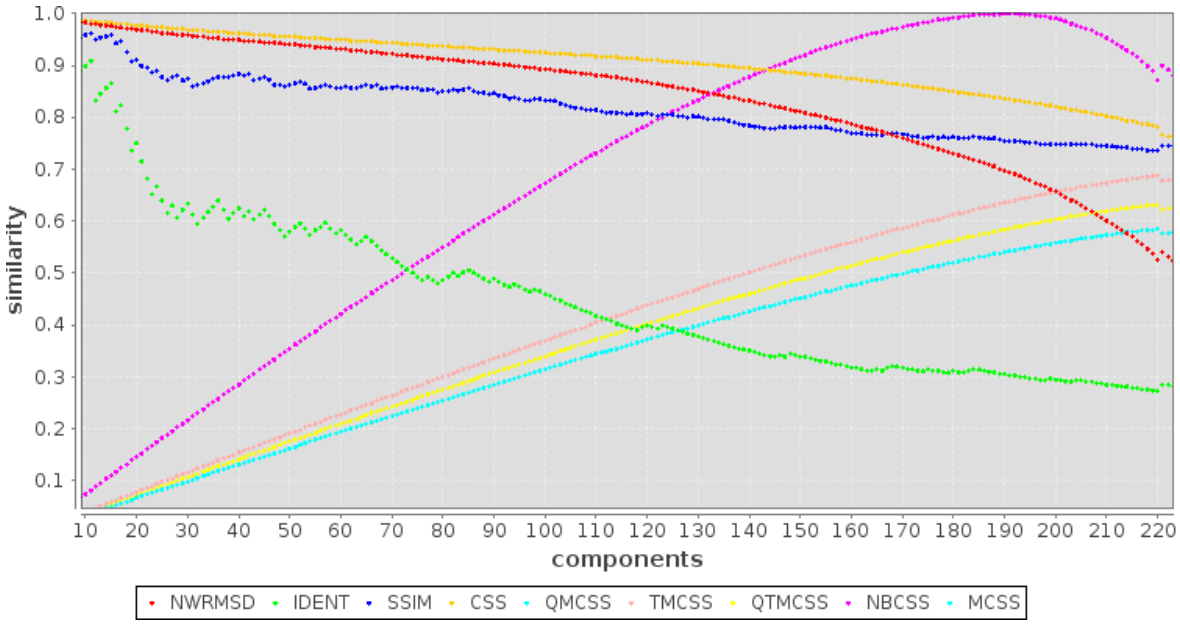


Abbildung 39: Deskriptoren des Matchs von 1Q1A und 1MA3.

Pfad, der über das gemeinsame Faltungsmotiv führt. Zhao et al. [188] charakterisiert das Epitop anhand von 20 Aminosäuren, die über die van-der-Waals- und/oder die Wasserstoffbrücken-Bindungen mit dem Ligand interagieren. Abb. 40 ist ein Auszug aus dem

COMPS	ATOMS	RMSD	NWRMSD	IDENT	SSIM	CSS	QMCSS	TMCSS	QTMCSS	NBCSS	MCSS	1	1	0.745				1	1	1	1	1	1	1	0,901	1	1	1	0.677	0.511	0.828
												A	G	T	F	R	Y	A	Q	I	H	V	F	G	T	S	V	N	L	Y	S
224	1461	2.453	0.511	0.281	0.745	0.76	0.579	0.681	0.626	0.864	0.681	A	G	A				A	Q	I	H	V	F	G	S	S	V	N	A	K	A
192	1237	1.853	0.691	0.302	0.755	0.833	0.544	0.64	0.588	1	0.64	A	G	A				A	Q	I	H	V	F	G	S	S	V	N	A	K	A
152	**955**	**1.411**	**0.807**	**0.336**	**0.781**	**0.883**	**0.457**	**0.537**	**0.494**	**0.924**	**0.537**	**A**	**G**	**A**				**A**	**Q**	**I**	**H**	**V**	**F**	**G**	**S**	**S**	**V**	**N**	**A**	**K**	**A**
128	803	1.192	0.854	0.383	0.8	0.905	0.394	0.463	0.426	0.825	0.463	A	G					A	Q	I	H	V	F	G	S	S	V	N	A	K	A
88	551	0.917	0.905	0.489	0.847	0.932	0.279	0.328	0.302	0.601	0.328	A	G					A	Q	I	H	V	F	G	S	S	V	N	A		A
87	546	0.913	0.906	0.494	0.849	0.933	0.276	0.325	0.298	0.594	0.325	A	G					A	Q	I	H	V	F	G	S	S	V	N			A
85	537	0.909	0.908	0.506	0.856	0.934	0.27	0.318	0.292	0.582	0.318	A	G						Q	I	H	V	F	G	S	S	V	N			A
55	340	0.694	0.936	0.582	0.858	0.953	0.178	0.21	0.193	0.388	0.21	A	G						Q	I	H	V		G	S	S	V	N			A
51	317	0.664	0.94	0.588	0.865	0.955	0.166	0.195	0.179	0.362	0.195	A							Q	I	H	V		G	S	S	V	N			A
50	311	0.655	0.941	0.58	0.863	0.956	0.163	0.191	0.176	0.355	0.191	A							Q	I	H	V		G	S	S		N			A
45	282	0.631	0.945	0.622	0.877	0.959	0.147	0.173	0.159	0.321	0.173	A							Q	I	H	V		G	S	S					A
41	256	0.605	0.948	0.61	0.881	0.961	0.134	0.158	0.145	0.293	0.158	A							Q	I	H	V		G	S						A
38	238	0.579	0.95	0.605	0.878	0.963	0.125	0.146	0.135	0.272	0.146	A							Q	I	H			G	S						A
35	219	0.554	0.953	0.629	0.874	0.966	0.115	0.135	0.124	0.252	0.135	A							Q	I	H			G							A
27	160	0.468	0.961	0.63	0.877	0.972	0.089	0.105	0.096	0.196	0.105	A								I	H			G							A
15	94	0.333	0.976	0.867	0.96	0.982	0.05	0.059	0.054	0.11	0.059	A								I	H										A
12	73	0.284	0.98	0.833	0.95	0.984	0.04	0.047	0.043	0.089	0.047	A								I											A
11	68	0.27	0.981	0.909	0.962	0.985	0.037	0.043	0.04	0.081	0.043	A								I											
6	44	0.187	0.987	0.833	0.932	0.989	0.02	0.024	0.022	0.045	0.024									I											

Abbildung 40: Integration des Epitops in den Ähnlichkeitskern ($COMPS = 152$).

Pfad der gemeinsamen Substrukturen. Zu Beginn erscheint das ILE des Epitops in einer CS aus 6 Aminosäuren ($COMPS = 6$). Zuletzt erscheint die Substitution TYR,ALA in einer CS aus 152 Aminosäuren. Da die Loop-Region der Apo-Struktur nicht kristallisiert worden

ist, konnten die Loop-Residuen der Holo-Struktur (PHE, ARG und TYR) nicht zugeordnet werden. Anhand des Verlaufs des NBCSS kann man feststellen, dass die beiden Strukturen über ein gemeinsames Faltungsmotiv verfügen, das unterhalb der BCS mit $NBCSS = 1.0$ zu finden ist. Darüber existierenden gemeinsamen Substrukturen, abschließend mit der MCS ($COMPS = 224$), sind das Ergebnis des geometrischen Rauschens der Residuenpaare, die nicht zum Kern der Ähnlichkeit der beiden Deacetylasen gehören. Die CS mit 152 Residuenpaaren kann als Kern der Ähnlichkeit der beiden Strukturen betrachtet werden: die $RMSD = 1.411\mathring{A}$ liegt unterhalb der Zwielichtzone (Abb. 10); die Ähnlichkeit der gemeinsamen Substruktur liegt bei $\approx 88.3\%$ ($CSS = 0.883$). Gemessen an den Größen der beiden Strukturen macht das Faltungsmotiv $\approx 49.4\%$ ($QTMCSS = 0.494$) der beiden Deacetylasen aus. Diese Analyse macht deutlich, dass 17 der 20 charakterisierten Aminosäuren zum starren, und 3 übrigen zum flexiblen Anteil des Epitops gehören. Die Identität des Kerns liegt bei $\approx 33.6\%$ ($IDENT = 0.336$). Die Identität des starren Anteils des Epitops liegt mit 5 Mutationen bei $^{12}/_{17} \approx 70.6\%$. Abgesehen von der TYR,LYS- und LEU,ALA-Substitution mit $\approx 51.1\%$ bzw. $\approx 67.7\%$, sind die Substitionen sehr ähnlich. Das Verhältnis der Identität des Kerns zu der Identität des starren Anteils des Epitops $^{70.6}/_{33.6} \approx 2.1$ zeigt, dass das Epitop bzw. die Funktion mehr als doppelt so hoch konserviert ist, als die Leitstruktur bzw. das Faltungsmotiv. Die Erkennung des Faltungsmotivs inklusive des Epitops, d.h. der Vergleich der Strukturen 1Q1A und 1MA3 dauert $\approx 8.5s$.

Abb. 38 zeigt auch die Matching-Ergebnisse für die 20 Aminosäuren nach Zhao et al. [188] (Abb. 38b) und $5\mathring{A}$-Umgebung (EpitopeMatch, Standardeinstellung) von OAD (Abb. 38c). Die in der Publikation Jakuschev and Hoffmann [77] verwendeten Matching-Modi Backbone/Specific (Abb. 33) bzw. Backbone/Unspecific (Abb. 31) sind Teilmengen des Modus AllAtoms/B62sAhMwNc7 (Standardeinstellung (EpitopeMatch)) und sind jeweils weniger aussagekräftig. Während EpitopeMatch zum Zeitpunkt der Publikation $\approx 138s$ bzw. $\approx 312s$ (auf einem Kern des Intel Core 2 Quad Q6600) für die Suche nach dem 20-Aminosäuren-Epitop brauchte, findet ihn die neue Version innerhalb von $\approx 1.2s$ (auf einem Kern des Xeon X5650). Eine bis zu 130-fache Performance-Steigerung bei einem deutlich umfangreicheren Analyse-Ergebnis.

Die PDB enthält gegenwärtig 7 Holo-Strukturen, mit einer oder mehreren Ketten, an die das OAD gebunden ist (Abb. 41, query/target structures). Ein weiteres Analyse-Werkzeug von EpitopeMatch ist der Kreuzvergleich. Abb. 41 zeigt vier Deskriptoren. Die Matrizen sind symmetrisch und die Anordnung der Epitop-Paare sind in allen vier Matrizen identisch. Die Größen der $5\mathring{A}$-Epitope verteilen sich zwischen 24 und 35 Residuen (Abb. 41, components counts, Diagonale). Besonders interessant ist der Vergleich der Epitope 1m2n_B_2001-OAD_B (35 Aminosäuren) und 4bv2_B_1248-OAD_B (28 Aminosäuren). Die entsprechenden Zellen sind mit dem Kreuz markiert. Beide Epitope haben 27 gemeinsame Residuen. Die $RMSD = 1.07\mathring{A}$ zeugt von einer sehr hohen geometrischen Ähnlichkeit. Die Identität $IDENT = 0.445$ ist hingegen relativ niedrig. Die Substitutionsähnlichkeit $SSIM = 0.821$ zeigt, dass die mutierten Aminosäuren durch sehr ähnliche substituiert sind. Die größenunabhängige Ähnlichkeit der gemeinsamen Substrukturen $CSS = 0.945$ ist sehr hoch. Die größenabhängige Ähnlichkeit $MCSS = 0.911$ liegt ebenfalls hoch. Alle 10 diskontinuierliche Epitope der 7 Strukturen sind einander sehr ähnlich ($CSS_{min} = 0.887$).

Die Analyse der homologen Strukturen und ihrer Epitope führt in beiden Fällen (2DN2 und 1Q1A & 1MA3) zu der Aussage, dass die unmittelbare Umgebung der Liganden (Epitope) physiko-chemisch und geometrisch deutlich höher konserviert ist als der Rest der Strukturen, wobei die Epitope ein integraler Bestandteil des Ähnlichkeitskerns (BCS) sind, der in der Regel ein gemeinsames stabiles Faltungsmuster ist.

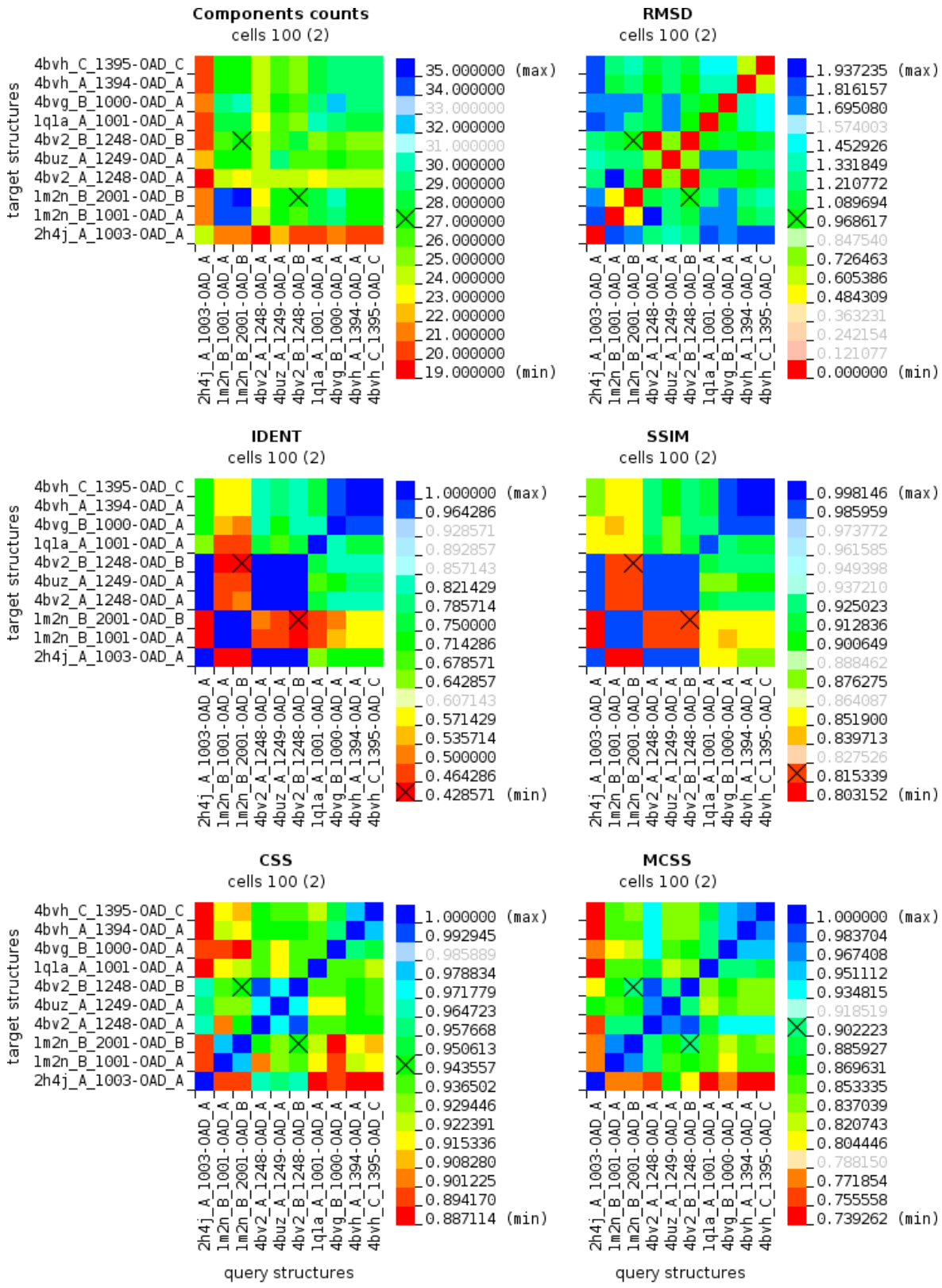


Abbildung 41: Kreuzvergleich der OAD-Epitope. Trotz einer starken Mutation (Epitope 1m2n_B_2001-OAD_B und 4bv2_B_1248-OAD_B, *IDENT* = 0.445 bei 27 korrespondierenden Residuen) bleibt die Funktion (Deacetylase) erhalten.

2.3.1.3 *Qualitativen Vergleiche - apo vs. holo*

Die Einschätzung der Flexibilität der Bindungsstellen ist wichtig für die Vorhersage der Proteinfunktion und für die Simulation des Protein-Ligand-Docking [57]. Die Autoren der

Publikation stellen sich die Frage, wie unterschiedlich sind die flexiblen und starren Bindungsstellen und anhand von welchen sequentiellen und strukturellen Eigenschaften sie differenziert werden können? Sie konzentrieren sich insbesondere auf: das Vorkommen der Residuen in den Bindungsstellen; die Interaktionen der Bindungsstellenresiduen mit den anderen Residuen; die Frequenz und die Verteilung der Konformationsänderungen im Bezug auf die Diederwinkel des Rückgrats; die unerlaubten Konformationen; die Architektur der Bindungsstellen-Loops. Die so ermittelten qualitativen Merkmale sollen im Rahmen der Vorhersage der Protein-Ligand-Komplexe und der templatebasierten Vorhersage der Proteinfunktion eingesetzt werden. Der untersuchte Datensatz besteht aus 98 Holo-Apo-Strukturpaaren. Alle Strukturpaare besitzen identische Primärstrukturen, womit die Untersuchung der Bindungsmechanismen der homologen Strukturen außen vor gelassen wird. Die Autoren beschäftigen sich also nur mit den Strukturpaaren, von denen es bekannt ist, dass sie über dieselben Bindungsmechanismen verfügen. Genau an diesem Punkt setzt EpitopeMatch an. Bevor die qualitativen Eigenschaften der Bindungsmechanismen abgeleitet werden können, müssen die ähnlichen Bindungsstellen über die quantitativen Untersuchungen einander zugeordnet werden. Erst wenn alle bekannten Bindungsstellen nach ihrer Ähnlichkeit geclustert sind, lassen sich Fragen beantworten wie: was ist der essentieller Bindungsmechanismus für einen bestimmten Ligand; welche Mutationen der Bindungsstelle führen zum Verlust der Bindungsfunktion oder zum Übergang in eine neue; welche Struktur ist für die Transplantation eines Epitops geeignet. Dieser Abschnitt behandelt vorerst die qualitativen Mustererkennungsmerkmale von EpitopeMatch und demonstriert die Eignung des MCSS für die Messung der RGD bzw. der FLX.

Jedes Strukturpaar aus dem Datensatz [57] wird nach der relativen Verschiebung der Cα-Atome der entsprechenden Bindungsstelle einer der drei Klassen zugeordnet: K_1 mit $\text{diff}(C\alpha) < 0.5\mathring{A}$ (keine Konformationsänderung); K_2 mit $0.5\mathring{A} \leq \text{diff}(C\alpha) \leq 2.0\mathring{A}$ (mittlere Konformationsänderung); und K_3 mit $\text{diff}(C\alpha) > 2.0\mathring{A}$ (große Konformationsänderung), mit $\text{diff}(C\alpha)$ als Distanzen zwischen einander entsprechenden Cα-Atomen der Bindungsstellen der überlagerten Holo-Apo-Strukturpaare. Die Residuen der Bindungsstellen sind

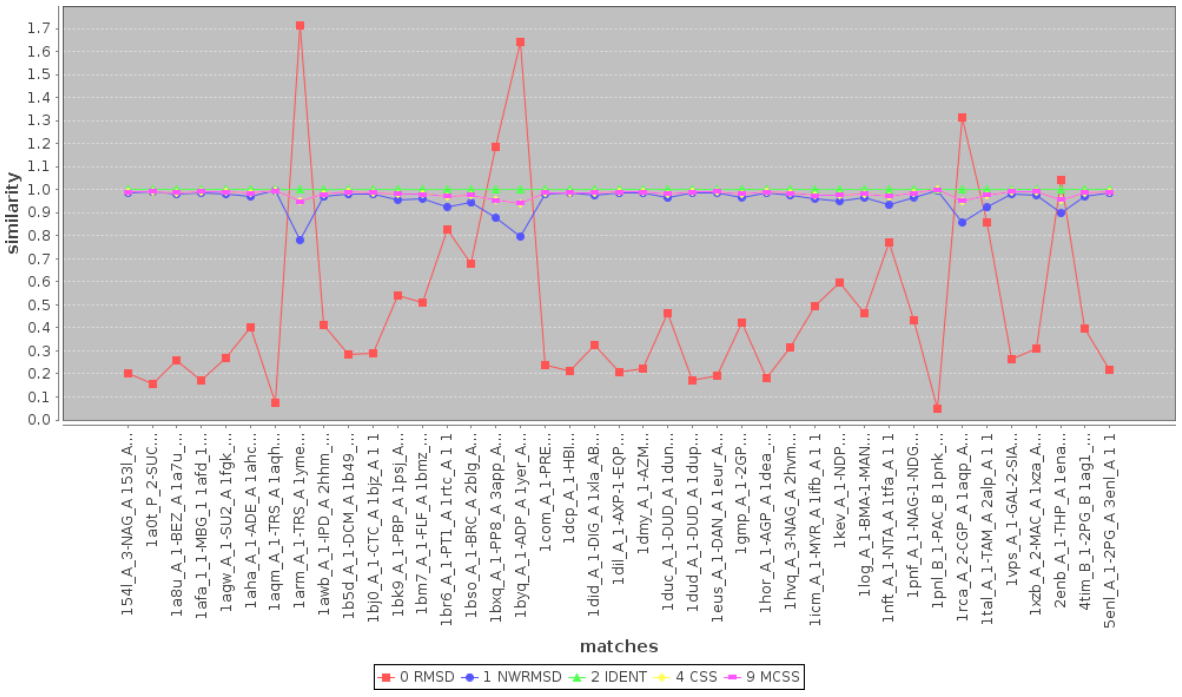


Abbildung 42: Klasse 1 (K_1 mit $\text{diff}(C\alpha) < 0.5\mathring{A}$).

alle Residuen, die mit mindestens einem Atom in die 4Å-Umgebung des Liganden hineinragen. Anhand der definierten Klassen der Bindungsstellen soll nun gezeigt werden, ob die Qualität der entwickelten Deskriptoren (Abs. 2.2.6.1) für die Erkennung und die Bewertung der Ähnlichkeit der Epitope und ihre quantitative Klassifizierung geeignet ist. Ferner soll diskutiert werden, ob EpitopeMatch alle Arten der Induced-Fit-Bewegungen erkennen kann. Abb. 42 zeigt die erste Klasse mit 41 Holo-Apo-Strukturpaaren. Abweichend zu Gunasekaran and Nussinov [57] sind Epitope als 5Å-Umgebungen der Liganden definiert. Dies schließt auch die Residuen ein, die nicht unmittelbar mit dem Liganden wechselwirken, aber zum Gerüst des Epitops gehören. Epitope aller drei Klassen sind diskontinuierlich. Die Epitope sind nicht allein nach den Cα-Atomen verglichen worden, sondern im Modus AllAtoms/B62sAhMwNc7 gegen die entsprechenden Apo-Strukturen gematcht. Die x-Achse enthält die Strutktur- und Chain-IDs der Holo-Strukturen, gefolgt von Ligand-IDs, Strutktur- und Chain-IDs der entsprechenden Apo-Strukturen und schließlich Match- und Permutation-Nummer. Die RMSD und das MCSS der Epitope der ersten Klasse

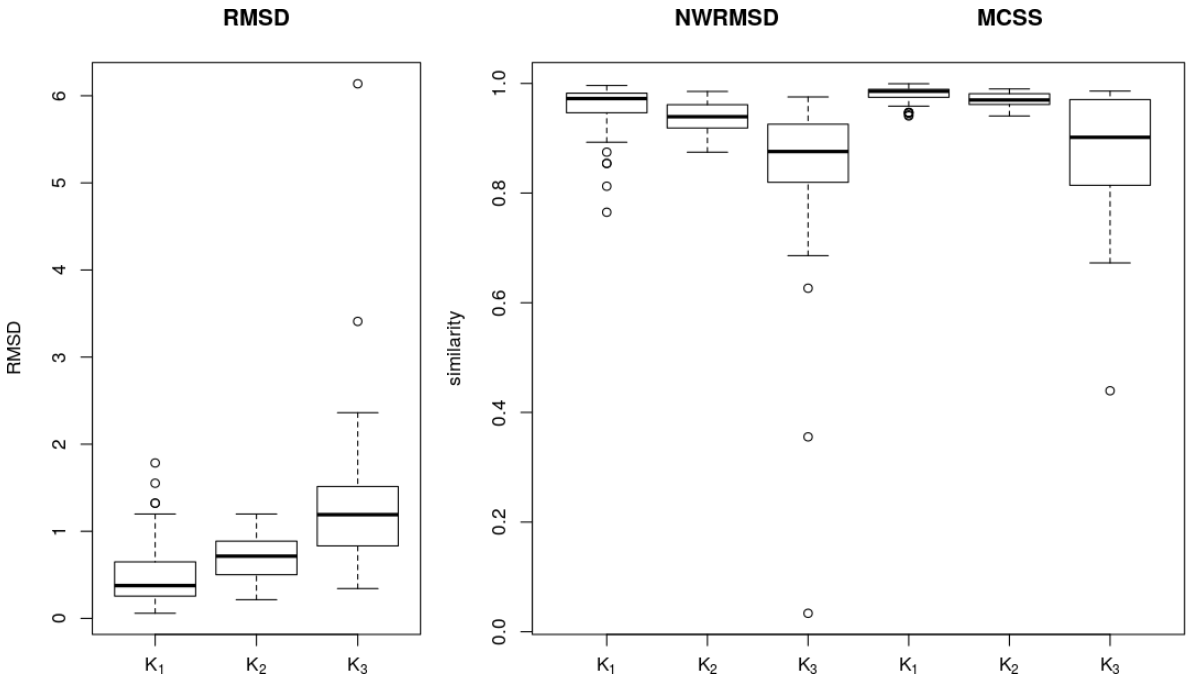
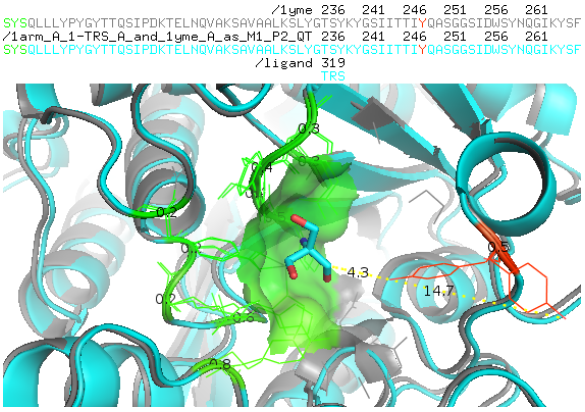


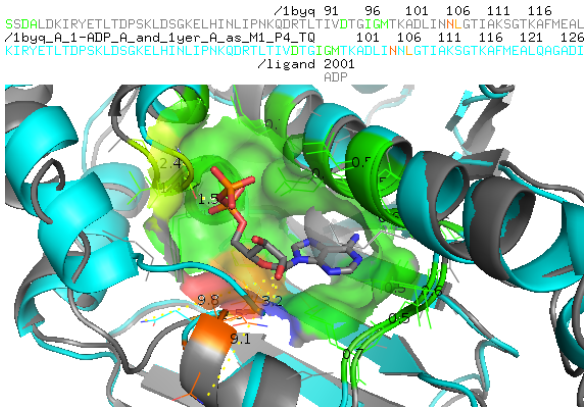
Abbildung 43: Deskriptoren der Klassen K_1 - K_3.

(Abb. 43) verteilen sich nach ihrer Transformation auf die Apo-Strukturen in den Bereichen $0.0588Å \leq RMSD(K_1) \leq 1.785Å$ und $94.07\% \leq MCSS(K_1) \leq 99.94\%$. Die meisten Matches liegen unterhalb $RMSD \leq 0.5Å$ ($Median = 0.3767Å$). Das Gesamtbild entspricht also größtenteils der Cα-Klassifizierung der ersten Klasse K_1 [57]. Das normalisierte RMSD-Äquivalent NWRMSD zeugt in der überwiegenden Anzahl der Fälle mit ($1st\,Qu. = 97.46\%$) von einer sehr guten geometrischen Übereinstimmung der Holo-Apo-Epitope. In den Fällen der Ausreißer sorgen die flexiblen Anteile der Epitope für die schlechteren Werte. Das CSS wird von MCSS in jedem einzelnen Fall überlagert ($MCSS = CSS$, Abb. 42), was bedeutet, dass alle Epitope vollständig erkannt worden sind. Da die Identität in allen Fällen $IDENT = 1.0$ ist, sind die physiko-chemischen Unterschiede gleich Null. Somit ist die gemessene Ähnlichkeit rein geometrischer Natur und kann der Rigidität gleich gesetzt werden ($MCSS = RGD$). Je größer der starre Anteil des Epitops, desto weniger fallen die Konformationsänderungen einzelner Residuen ins Gewicht. So ist die Ähnlichkeit des größten RMSD-Ausreißers 1AR-

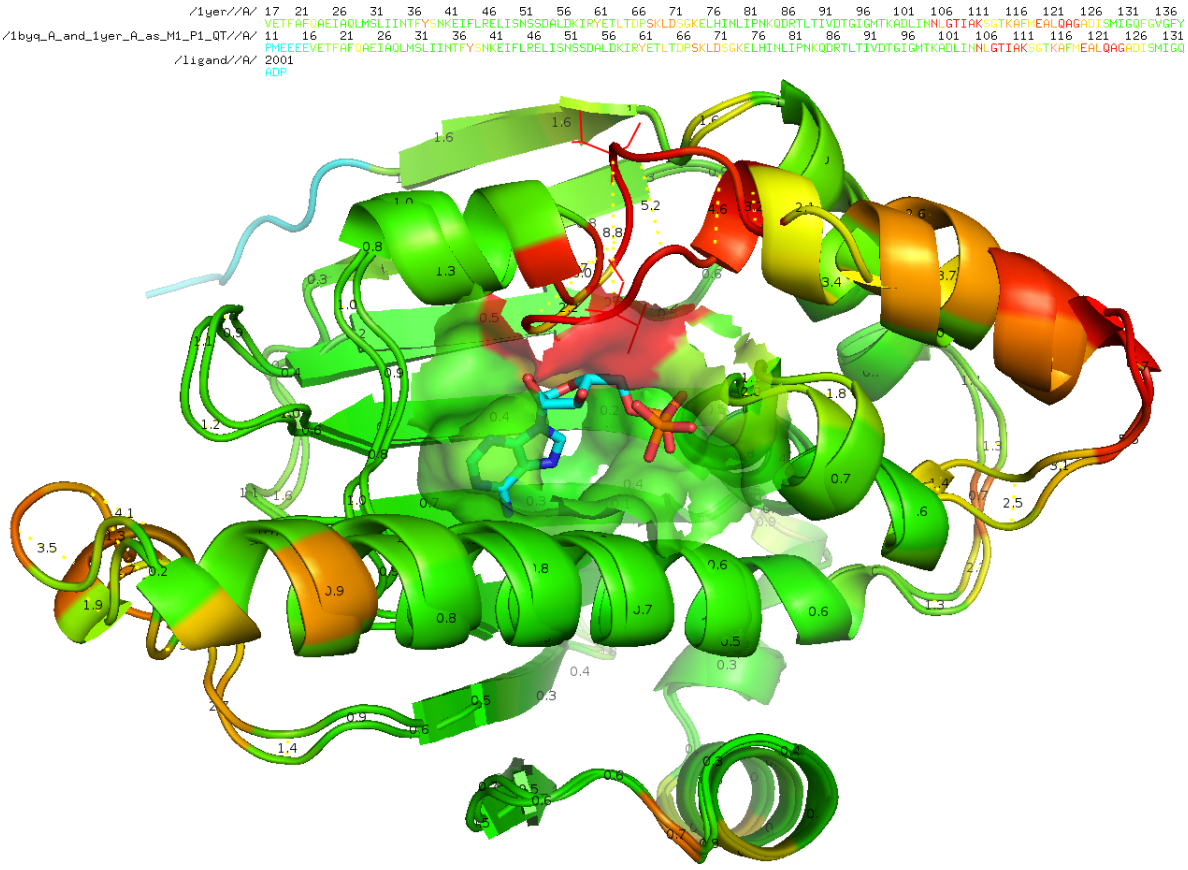
M/1YME $MCSS = 94.6\%$ mit dem größeren starren Anteil von $^{10}/_{11}$ Residuen (Abb. 44a) etwas besser als die Ähnlichkeit des 1BXQ/3APP-Epitops $MCSS_{min} = 94.07\%$ mit dem kleineren starren Anteil von $^{21}/_{27}$ Residuen. Bei der näheren Betrachtung des größten Aus-



(a) Query(1ARM,A,TRS) / Target(1YME,A).



(b) Holo-Epitop(1BYG,A,ADP) / Apo-Struktur(1YER,A).



(c) Holo-Struktur(1BYG,A,ADP) / Apo-Struktur(1YER,A).

Abbildung 44: Beispiele aus der Klasse 1 nach [57]. Beide Beispiele demonstrieren, dass ihre Einordnung in die Klasse "keine Konformationsänderung" falsch ist. Im Fall (a) findet eine Rotamer-Bewegung (rotamer motion) und im Fall (b) bzw. (c) eine Loop-Bewegung (loop motion) statt. Die Färbung erfolgt nach der $NWRMSD(RT_{qt}^{QT})$ (min max) der korrespondierenden Residuen.

reißers (Abb. 44a) stellt man fest, dass ein Rotamer (TYR248) für die deutliche Verschlechterung der RMSD verantwortlich ist. Mit 11 Residuen ist das Epitop relativ klein, sodass das

Verhältnis der Anzahl der stark deplatzierten TYR248-Restatome zu der Anzahl aller Atome des Epitops einen spürbar negativen Einfluss auf die RMSD ausübt. Das TYR248 muss mit der kürzesten Residuum-Ligand-Distanz von $\approx 4.3\mathring{A}$ als ein nicht zum Epitop gehörendes Residuum betrachtet worden sein, da er sich außerhalb der $4\mathring{A}$-Umgebung befindet. Abb. 44a zeigt jedoch deutlich, dass TYR248 eindeutig ein Teil des Bindungsmechanismus ist. Das $4\mathring{A}$-Umgebung-Threshold ist in diesem Fall einfach zu strikt. Ein Grund mehr, der für die Standardeinstellung von EpitopeMatch von $5\mathring{A}$-Umgebung-Threshold spricht. Die maximale Verschiebung der Atome des TYR248-Rests beträgt in diesem Fall $\approx 14.7\mathring{A}$ bei der gleichzeitigen Cα-Verschiebung des TYR248 von lediglich $\approx 0.5\mathring{A}$. Dies zeigt, dass die alleinige Betrachtung der Cα-Verschiebungen für die Klassifizierung nicht ausreichend ist und der Einsatz des AllAtoms-Templates gerechtfertigt ist. Die Art des Induced-Fit ist in diesem Fall *rotamer motion*. Der zweitgrößte RMSD-Ausreißer (Abb. 44b/Abb. 44c), ein Epitop mit 18 Residuen in der $5\mathring{A}$-Umgebung des ADP, offenbart eine Bewegung des Loops. Die Transformation des Epitops ist in der Darstellung (Abb. 44b) vertauscht, sodass das Query in grau und das Target in cyan dargestellt wird. Die Oberflächendarstellung entspricht somit der Oberfläche der QS. Es ist gut zu erkennen, dass zwischen zwei helikalen Abschnitten ein Loop integriert ist (links, unten). Im Holo-Zustand (grau) rückt das ASN106 mit $3.2\mathring{A}$ an das ADP und wird zum Bestandteil der Helix (links, unten, grau). Die restlichen Residuen des Loops sind dabei von der Bindungsstelle weit entfernt und die Helix (rechts, oben, grau) ist entwunden. Im Apo-Zustand (cyan) nimmt der Abschnitt (rechts, oben, cyan) wieder die helikale Form an, der Loop (cyan) bewegt sich in die Richtung der Bindungsstelle und ASN106 entfernt sich von dem ADP auf eine Distanz von $9.8\mathring{A}$. In diesem Fall ist eine Ungenauigkeit von EpitopeMatch zu erkennen: das ASN106 der QS wird dem ASN105 der TS zugeordnet. Der Algorithmus hat sich in diesem Fall für das näher liegende ($9.1\mathring{A}$) Residuum entschieden. Mit ASN105 liegt eine Kombination mit besseren Deskriptorwerten vor, als mit dem weiter entfernten ASN106. Die Kombinationsbildung ist jedoch über die Algorithmusparameter steuerbar. Dieses Defizit kann durch die Erhöhung der Anzahl der Permutationen $BCST_{min}$ pro Ordnung k (Abb. 15) überwunden werden, sodass die richtig-positive Zuordnung ASN106-ASN106 in einer der schlechter bewerteten Permutationen auftauchen würde. Die Cα-Verschiebung des ASN106 liegt bei $\approx 3.5\mathring{A}$ (verdeckt von der Helix). Abb. 44c zeigt die beiden vollständig überlagerten Strukturen. Die QS ist wieder cyan. Die grün gefärbten Residuenpaare repräsentieren den starren Kern. Der Rest des Farbspektrums, angefangen mit gelb über orange bis rot zeigt die flexiblen gemeinsamen Substrukturen. Die für die Lösung zugänglichen Oberflächenbereiche in der unmittelbaren Umgebung des Liganden beziehen sich hier auf die Apo-Struktur. Die rot schattierten Oberflächenbereiche gehören zu den Aminosäuren des Loops. Obwohl der Ligand nach der Transformation der Holo-Struktur auf die Apo-Struktur bis auf $1.3\mathring{A}$ an das Loop-Residuum THR109 heran kommt, ist es denkbar, dass er auch in der Apo-Position des Loops gebunden sein kann, da der Loop die Bindungsstelle nicht verschließt, sondern eher ergänzt und somit eindeutig ein Teil des Bindungsmechanismus ist. Die dominierende Art des Induced-Fit ist in diesem Fall *loop motion*. Dieses Holo-Apo-Strukturpaar ist in der Klasse K_1 (keine Konformationsänderung) eindeutig fehl am Platz.

Abb. 45 zeigt die zweite Klasse mit 35 Holo-Apo-Strukturpaaren. Auch hier sind alle Epitope vollständig erkannt ($MCSS = CSS$). Die mittlere RMSD der Epitope (Abb. 43) mit $Median = 0.7145\mathring{A}$ ist höher als die mittlere RMSD der ersten Klasse. Allerdings liegen alle Matches weit unterhalb der Zwielichtzone (Abb. 10). Mit $MCSS_{min} = 94.07\%$ können alle Holo-Apo-Epitope als sehr ähnlich angesehen werden. Der Trend der wachsenden RMSD- und sinkenden NWRMSD- und MCSS-Werte (Abb. 43) zeugt von einer steigenden Flexibilität der Epitope.

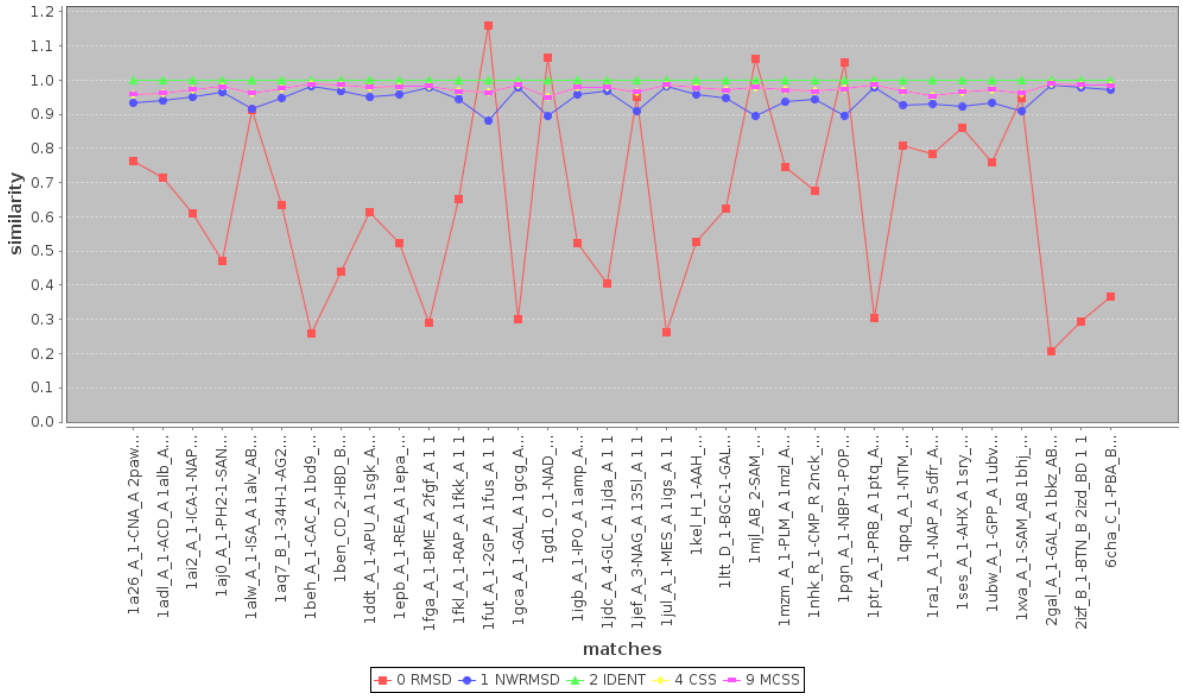


Abbildung 45: Klasse 2 (K_2 mit $0.5\mathring{A} \leq \text{diff}(C\alpha) \leq 2.0\mathring{A}$).

Abb. 46 zeigt die dritte Klasse mit 22 Holo-Apo-Strukturpaaren. Erstmalig erscheinen

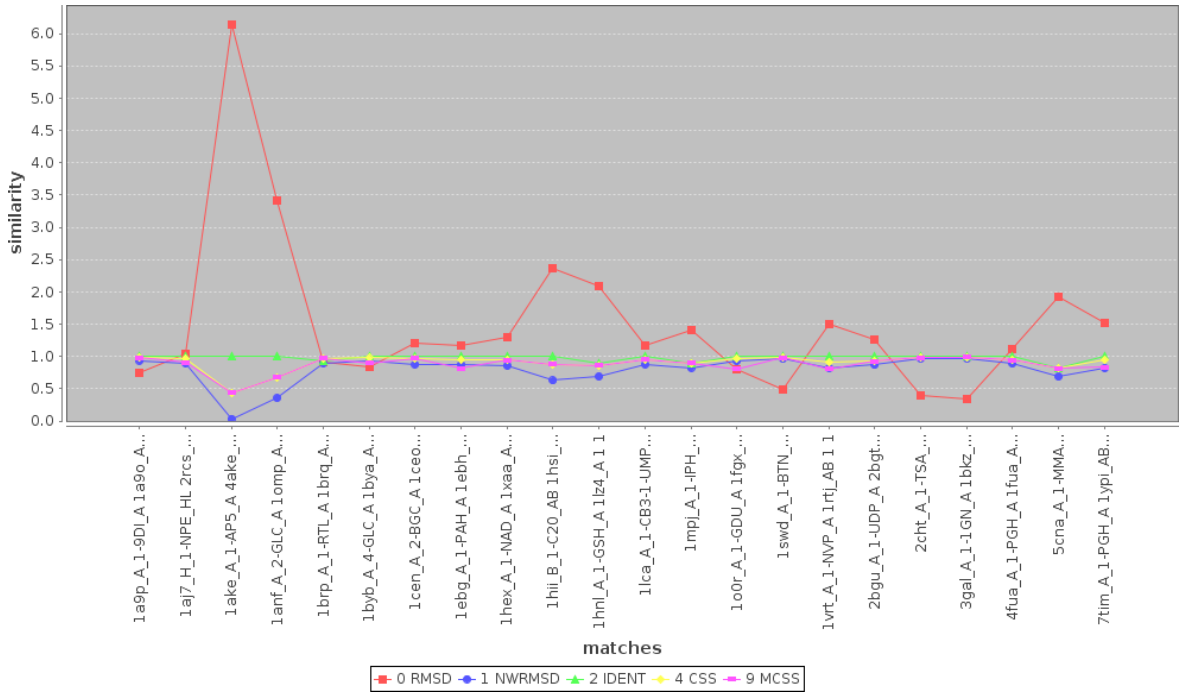
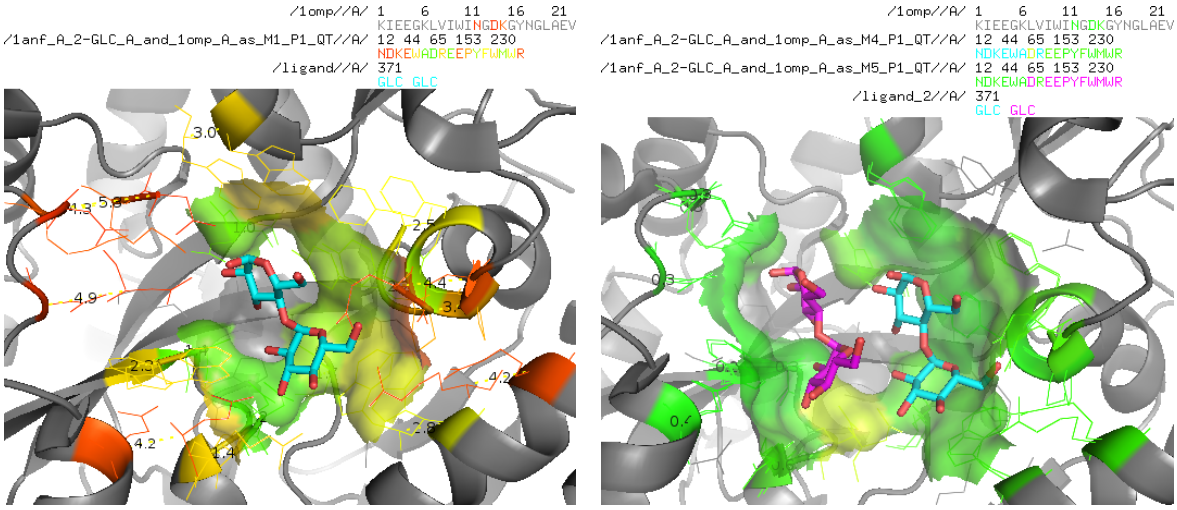


Abbildung 46: Klasse 3 (K_3 mit $\text{diff}(C\alpha) > 2.0\mathring{A}$).

Matches mit unvollständig erkannten Holo-Epitopen ($MCSS \neq CSS$ und/oder $IDENT \neq 1.0$). Existiert nur ein unvollständiger Match (1BYB/1BYA, 1EBG/1EBH, 1HNL/1HSL, 1OoR/1FGX), dann handelt es sich meist um eine sehr große Konformationsänderung, die durch die Bewegung eines Loops (*loop motion*) hervorgerufen worden ist. Der starre Anteil des Epitops ist in diesen Fällen viel größer als der flexible Anteil, wobei das Match teilwei-

se die Residuen des flexiblen Anteils enthalten kann. Die Zuordnung der richtig-positiven Residuen des flexiblen Anteils des Epitops ist in solchen Fällen bedingt durch ein starkes geometrisches Rauschen der falsch-positiven Residuen und die fehlende Primärstrukturinformation der diskontinuierlichen Epitope nicht möglich. Durch die Wahl einer strikteren Substitutionsmatrix (Abb. 33) kann man die Zuordnung der richtig-positiven Residuen des flexiblen Anteils erzwingen, allerdings würde man durch ihre Wahl das Matchen der homologen Strukturen ausschließen. Existieren mehrere unvollständige Matches (1AKE/4AKE, 1ANF/1OMP), dann handelt es sich um eine Konformationsänderung als Resultat der Domänenbewegung der Struktur (*domain motion*). Die so genannten Hinges (Abs. 2.2.7.1), d.h. für die Domänenbewegung verantwortlichen Residuen, können durchaus ein Bestandteil der Epitope selbst sein. So wie im Fall des Holo-Apo-Strukturpaares 1ANF/1OMP, das im Hinblick auf die Domänenbewegung im Abs. 2.2.7 ausführlich untersucht wird. Bei dem zweitgrößten RMSD-Ausreißer der Klasse K_3 (Abb. 43) handelt es sich um das aus 17 Residuen bestehende Epitop des 1ANF/1OMP (Abb. 47). GLU111, eins der Residuen des



(a) starr (b) flexibel

Abbildung 47: 1ANF/1OMP-Epitop, starres & flexibles Matchen.

Epitops, gehört zu einer der drei Hinge-Bending-Regionen (Abs. 2.2.7) der Struktur. Das vollständig bzw. starr gematchte Epitop (Abb. 47a) befindet sich mit $RMSD = 3.4017\mathring{A}$ in der Zwielichtzone (Abb. 10), bekommt jedoch mit $MCSS = 64.99\%$ eine relativ hohe Ähnlichkeitswertung. Dies hängt mit der Beschaffenheit des Epitops zusammen, der aus zwei starren Teilen besteht, die einzeln betrachtet ihre Konformation nur geringfügig ändern (Abb. 47b). Die Deskriptoren des flexibel gematchten Epitops

$$DRMSD(CS_{10}, CS_7) = \frac{10 \cdot 1.1872\mathring{A} + 7 \cdot 0.6187\mathring{A}}{17} = 0.9531\mathring{A}$$

$$DMCSS(CS_{10}, CS_7) = 0.5464 + 0.3996 = 0.946 \triangleq 94.6\%$$

zeigen, das die Residuenkonformation im Holo- und Apo-Zustand des Epitops nahezu unverändert bleibt. Der Induced-Fit-Anteil, der aufgrund der Domänenbewegung entsteht

$$RMSD(CS_{17}) - DRMSD(CS_{10}, CS_7) = 3.4017\mathring{A} - 0.9531\mathring{A} = 2.4486\mathring{A}$$

$$DMCSS(CS_{10}, CS_7) - MCSS(CS_{17}) = 0.946 - 0.6499 = 0.2961 \triangleq 29.61\%$$

ist allerdings relativ hoch.

Der größte RMSD-Ausreißer der Klasse K_3 (Abb. 43), ein Epitop aus 45 Aminosäuren, verteilt sich über drei Domänen der Strukturen 1AKE/4AKE. Das Match liegt mit $RMSD = 6.1382\mathring{A}$ weit oberhalb der Zwielichtzone und würde mit diesem Wert in der Regel als falschpositives Match durch den Raster fallen. Allerdings auch in diesem Fall deutet das relativ hohe $MCSS = 0.4395$ auf das Vorhandensein der gut konservierten starren Anteile. Das Epitop besteht aus drei starren Substrukturen. Die flexiblen Deskriptorwerte

$$DRMSD(CS_{20}, CS_{14}, CS_{11}) = \frac{20 \cdot 1.624\mathring{A} + 14 \cdot 1.538\mathring{A} + 11 \cdot 1.354\mathring{A}}{45} = 1.531\mathring{A}$$

$$DMCSS(CS_{20}, CS_{14}, CS_{11}) = 0.398 + 0.286 + 0.223 = 0.907 \triangleq 90.7\%$$

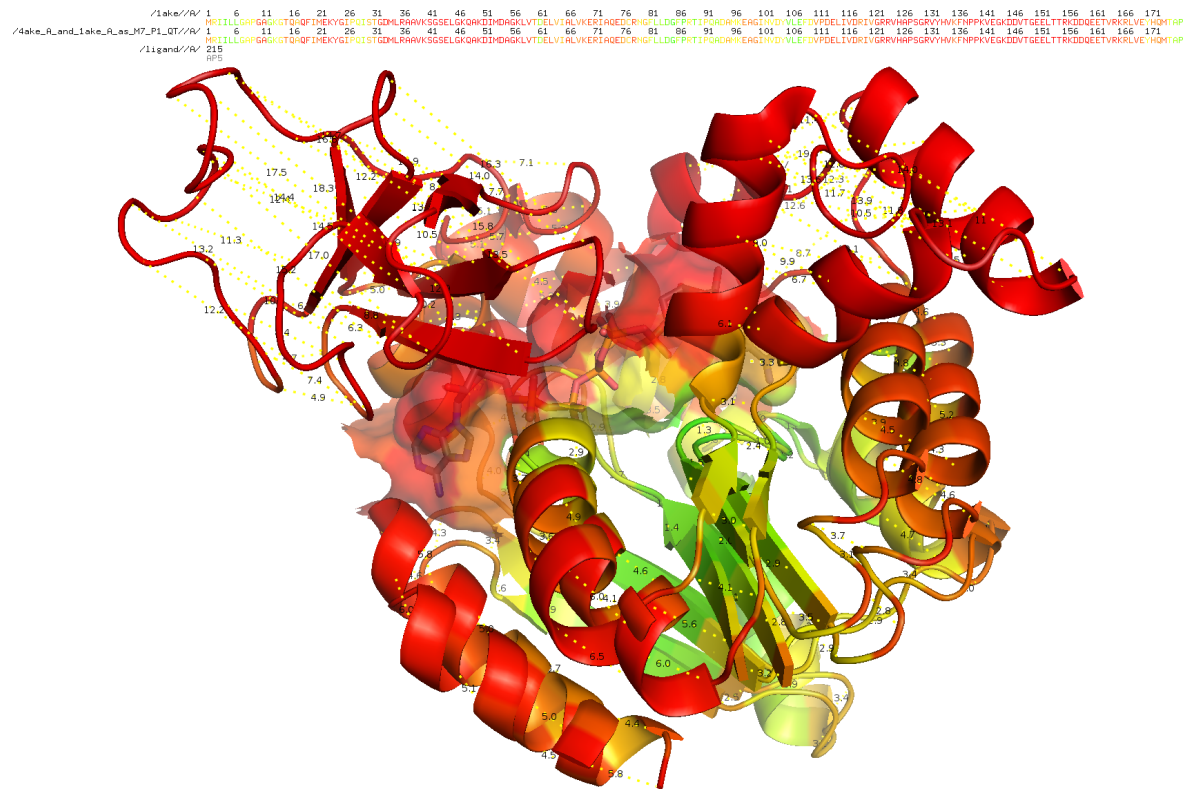
des richtig-positiven Matchs heben allerdings die Ähnlichkeit des Epitops im Holo- und Apo-Zustand unter Ausschluss der Domänenbewegung deutlich hervor. Abb. 48a demonstriert starr gematchte, vollständige Apo- und Holo-Struktur. Der Ligand ist im Holo-Zustand der Struktur im Inneren des Proteins verborgen. Die Oberflächendarstellung entspricht den Oberflächenanteilen der Holo-Struktur. Die $RMSD = 7.204\mathring{A}$ der insgesamt 214 Aminosäuren ist außerordentlich hoch. Der $MCSS = 0.531$ liegt mit 53.1% innerhalb des Ähnlichkeitsbereichs für homologe Strukturen $30.0\% < MCSS < 100.0\%$. Die Domänenbewegungen gehen mit den Verschiebungen der einzelnen Cα-Atome von bis zu $20\mathring{A}$ einher. Abb. 48b demonstriert flexibel gematchte Holo- und Apo-Struktur. Der Algorithmus erkennt 3 Domänen mit 125 (cyan), 51 (magenta) und 38 (blau) Residuentupeln. Die Holo-Struktur ist domänenweise dargestellt und nach der Ähnlichkeit der transformierten Residuen gefärbt (im Farbspektrum grün über gelb nach rot). Die Oberflächenanteile sind entsprechend der Domänenzugehörigkeit gefärbt. Die Strukturalignments sind in den beiden Fällen (starr, flexibel) identisch und resultieren in zwei Hinge-Bending-Regionen: Schnittstelle cyan/blau mit ILE29/SER30 und LEU67/VAL68; Schnittstelle cyan/magenta mit ILE116/VAL117 und ARG167/LEU168. EpitopeMatch identifiziert insgesamt 13 Aminosäuren (*dots*) als Hinge-Bending-Residuen. Interessante Feststellung ist, dass die Domänenübergänge in diesem Fall reich an hydrophoben Aminosäuren (ILE, LEU, VAL) sind. Womöglich liegt der Hinge-Bending-Mechanismus im für die Lösung unzugänglichen Bereich. Während der relativ große Ligand (AP5) im Holo-Zustand vollständig umschlossen ist, ist die Struktur im Apo-Zustand sehr weit geöffnet, sodass die Oberflächenanteile des Epitops für die Lösung zugänglich sind. Die flexiblen Deskriptorwerte

$$DRMSD(CS_{20}, CS_{14}, CS_{11}) = \frac{125 \cdot 1.532\mathring{A} + 51 \cdot 1.73\mathring{A} + 38 \cdot 1.999\mathring{A}}{214} = 1.662\mathring{A}$$

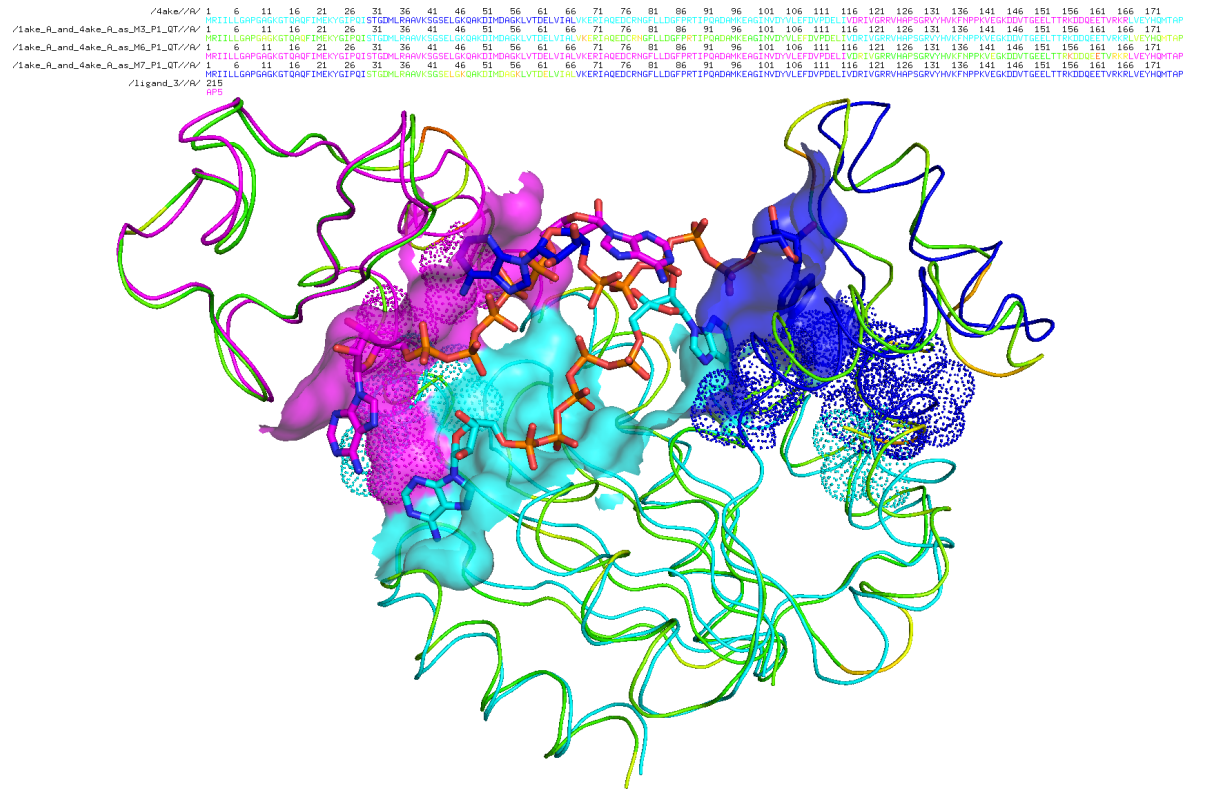
$$DMCSS(CS_{20}, CS_{14}, CS_{11}) = 0.525 + 0.22 + 0.151 = 0.907 \triangleq 90.7\%$$

zeigen, dass die Domänen ihre Faltungsmuster größtenteils beibehalten.

Tab. 12 fasst die qualitative Untersuchung der Holo-Apo-Epitope zusammen. Die Spalte *class* enthält die Nummern der drei Klassen aus [57]. Die Hintergrundfarben grün für keine, gelb für mittlere und rot für große Konformationsänderung entsprechen dem Verlauf des Farbspektrums, das in den PyMol-Abbildungen für die Darstellung der Ähnlichkeit der Residuenpaare verwendet wird. *QUERY* enthält die Bezeichnungen der Epitope, die sich aus der PDB-ID der Holo-Struktur, der Chain-ID(s) des Epitops und den Ligand-ID(s) und deren Anzahl zusammensetzt. *TARGET* enthält die PDB-IDs der Apo-Strukturen und deren Chain-ID(s). *COMPS* gibt die Größe des jeweiligen Epitops bzw. Matchs in Residuen an. *IDENT* stellt den Prozentsatz der Residuenpaare mit identischen Residuen dar. Da



(a) Starr gematchte Apo- (4AKE) und Holo-Struktur (1AKE). Die Oberflächendarstellung entspricht den Oberflächenanteilen der Holo-Struktur. Die Färbung erfolgt nach der $NWRMSD(RT_{qt}^{QT})$ der korrespondierenden Residuen. Die einzelnen Domänen sind nicht sichtbar.



(b) Flexibel gematchte 1AKE und 4AKE. Die Struktur teilt sich in 3 Domänen mit 125 (cyan), 51 (magenta) und 38 (blau) Residuen. Die Holo-Struktur ist domänenweise dargestellt und analog zu (a) gefärbt. Die Apo-Struktur und ihre Oberflächenanteile sind entsprechend der Domänenzugehörigkeit gefärbt. Die Verschiebung des Liganden (AP5) entspricht der Transformation der einzelnen Domänen. Die Hinge-Bending-Residuen sind als Punkte (*dots*) dargestellt.

Abbildung 48: 1AKE/4AKE, starres & flexibles Matchen. Der größte Ausreißer aus der Klasse 3.

alle Holo-Apo-Strukturpaare über identische Primärstrukturen verfügen, entspricht dieser Wert dem Anteil der richtig-positiv zugeordneten Residuenpaare pro Epitop. Die Spalte *rigid parts* enthält die Anzahl der starren Anteile pro Epitop. Die RMSD bezieht sich auf das gesamte Epitop gematcht im Modus AllAtoms/B62sAhMwNc7. Die DRMSD ist berechnet, falls ein Epitop sich aus mehreren starren Substrukturen zusammensetzt. Der MCSS drückt die Ähnlichkeit eines Epitops aus. Der DMCSS drückt die Ähnlichkeit eines Epitops unter Ausschluss der Domänenbewegung aus. Die Spalte *induced fit* bezeichnet die Art der dominanten Substrukturenbewegungen, die zu den Konformationsänderungen der Epitope geführt haben. Die Tabelle ist nach dem DMCSS aufsteigen sortiert. Die Matches der größten Klasse K_1 befinden sich überwiegend im oberen Drittel der Tabelle. Die Paare der kleineren Klasse K_2 verteilen sich verstärkt im mittleren Bereich. Den unteren Bereich der Tabelle nehmen die Paare der Klasse K_3 ein. Im Wesentlichen ist die Klassifizierung nach [57] zutreffend, allerdings enthält jede der drei Klassen mehrere Ausreißer, die in der definierten Klasse fehl am Platz sind. Die alleinige Betrachtung der Verschiebungen der Cα-Atome ist demnach für eine genaue Klassifizierung bei weitem nicht ausreichend. Aus der Sicht von EpitopeMatch beginnen die Substrukturenbewegungen, gemessen am Vergleich der Holo- und Apo-Konformationen der diskontinuierlichen Epitope, mit den Positionsänderungen der Residuenreste (*rotamer*) bei nahezu gleichbleibenden Positionen der dazugehörigen Cα-Atome (Abb. 44a) - die Faltung bleibt weitgehend erhalten. Die nächste Stufe des Induced-Fit sind die Konformationsänderungen der kleineren Sekundärstrukturabschnitte wie der Loops (*loop*) und der α-Helices bzw. β-Faltblätter (*backbone*), die im Vergleich zum starr bleibenden Anteil der gesamten Struktur zu klein sind, um als Domänen bezeichnet zu werden (Abb. 44c) - die Faltung verändert sich nur innerhalb der kleineren Abschnitte. Die letzte Stufe des Induced-Fit ist die Domänenbewegung (Abb. 24, Abb. 48a) - die Faltung der Domänen bleibt erhalten, die wesentlichen Konformationsänderungen finden in den Hinge-Bending-Regionen statt. Die Stufen zwei und drei können selbstverständlich die untergeordneten Stufen inkludieren. Nach EpitopeMatch wird die Klassifizierung wie folgt ausgelegt: Matches mit $0.0\mathring{A} \leq RMSD < 0.5\mathring{A}$ sind Substrukturen, die abzüglich der möglichen Fehler der Strukturbestimmungsmethoden praktisch identisch sind. Im Bereich $0.5\mathring{A} \leq RMSD < 1.0\mathring{A}$ werden die kleineren Konformationsänderungen spürbar. Im Bereich $1.0\mathring{A} \leq RMSD < 1.5\mathring{A}$ steigt das Verhältnis der flexiblen zu den starren Substrukturen der Epitope stetig an. Allerdings werden in der Regel alle Epitope im Bereich $0.0\mathring{A} \leq RMSD < 1.5\mathring{A}$ vollständig erkannt, sodass er als eine einzige Klasse K_1^{EM} zusammengefasst werden kann. Diese Klasse entspricht der oben beschriebenen Stufe 1 des Induced-Fit, mit einem großen starren und einem kleinen flexiblen Anteil, der eher geringfügigen Konformationsänderungen unterliegt. Die Ähnlichkeit bewegt sich dabei im Bereich $0.9 < MCSS \leq 1.0$. Zu dieser Klasse (Tab. 12, Spalte *class EM*, grün) gehören 85 der insgesamt 98 Holo-Apo-Epitoppaare. Im Bereich $1.5\mathring{A} \leq RMSD < 3.0\mathring{A}$, der sich bis zur unteren Grenze der Zwielichtzone erstreckt, tauchen die ersten falsch-positiven Residuenpaare auf. Die gemeinsamen Substrukturen dieser Klasse K_2^{EM} (Spalte *class EM*, gelb, 11 Paare) entsprechen der Stufe 2 des Induced-Fit und bestehen aus einem großen starren und einem kleinen flexiblen Anteil, der großen Konformationsänderungen unterliegt. Die Ähnlichkeit bewegt sich in diesen Fällen im durchschnittlichen Bereich $0.8 < MCSS \leq 9.0$. Zu der letzten Klasse K_3^{EM} (Spalte *class EM*, rot) gehören lediglich 2 Paare aus dem Datensatz. Die $RMSD \geq 3.0\mathring{A}$ bewegt sich in diesen Fällen in und oberhalb der Zwielichtzone. Die gemeinsamen Substrukturen bestehen dabei aus mehreren starren und flexiblen Anteilen, wobei die starren Substrukturen über die unterschiedlichen Domänen verteilt sind, die sich während des Induced-Fit relativ zueinander verschieben. Betrachtet man jedoch die Domänen-Deskriptoren (DRMSD, DMCSS) der Klasse K_3^{EM}, so liegen die Werte unter Aus-

class	QUERY	TARGET	COMPS	IDENT	RMSD	DRMSD	MCSS	DMCSS	induced fit	domains	class EM
1	1pnl_B_1-PAC_B	1pnk_A	12	1	0,0588		0,9994		none	1	1
1	1xzb_A_2-MAC_A	1xza_A	20	1	0,2749		0,9923		none	1	1
1	1eus_A_1-DAN_A	1eur_A	16	1	0,1980		0,9921		none	1	1
1	1afa_1_1-MBG_1	1afd_1	7	1	0,1950		0,9909		none	1	1
1	1aqm_A_1-TRS_A	1aqh_A	9	1	0,4137		0,9906		none	1	1
1	1aot_P_2-SUC_P	1aos_P	18	1	0,1705		0,9904		none	1	1
1	154l_A_3-NAG_A	153l_A	18	1	0,2057		0,9904		none	1	1
2	2gal_A_1-GAL_A	1bkz_AB	7	1	0,2150		0,9900		none	1	1
1	5enl_A_1-2PG_A	3enl_A	15	1	0,2127		0,9898		none	1	1
1	1hor_A_1-AGP_A	1dea_A	22	1	0,1932		0,9898		none	1	1
1	1b5d_A_1-DCM_A	1b49_A	15	1	0,2846		0,9893		none	1	1
1	1agw_A_1-SU2_A	1fgk_A	15	1	0,2641		0,9893		none	1	1
1	1dud_A_1-DUD_A	1dup_A	12	1	0,2872		0,9890		none	1	1
1	1dil_A_1-AXP-1-EQP_A	2sil_A	20	1	0,2163		0,9888		none	1	1
1	1dmy_A_1-AZM_A	1dmx_A	16	1	0,2411		0,9882		none	1	1
1	1dcp_A_1-HBI_A	1dco_A	9	1	0,2231		0,9872		none	1	1
1	1a8u_A_1-BEZ_A	1a7u_A	12	1	0,2575		0,9871		none	1	1
2	1gca_A_1-GAL_A	1gcg_A	18	1	0,3026		0,9870		none	1	1
2	2izf_B_1-BTN_B	2izd_BD	17	1	0,2988		0,9869		none	1	1
1	1bk9_A_1-PBP_A	1psj_A	15	1	0,5415		0,9868		none	1	1
1	4tim_B_1-2PG_B	1ag1_O	17	1	0,3405		0,9866		none	1	1
2	1jul_A_1-MES_A	1igs_A	13	1	0,3089		0,9863		none	1	1
2	1ptr_A_1-PRB_A	1ptq_A	11	1	0,3042		0,9862		none	1	1
1	1aha_A_1-ADE_A	1ahc_A	11	1	0,3767		0,9862		none	1	1
1	1vps_A_1-GAL-2-SIA_A	1vpn_A	19	1	0,3722		0,9859		none	1	1
2	1beh_A_1-CAC_A	1bd9_AB	14	1	0,2683		0,9858		none	1	1
1	1com_A_1-PRE_A	2chs_A	6	1	0,2560		0,9857		none	1	1
2	6cha_C_1-PBA_BCF	4cha_ABCEFG	17	1	0,3731		0,9855		none	1	1
2	3pca_MN_3-DHB_MNO	2pcd_ABCDEFMNOPQR	18	1	0,3006		0,9850		none	1	1
1	1bjo_A_1-CTC_A	1bjz_A	16	1	0,4415		0,9842		none	1	1
1	1hvq_A_3-NAG_A	2hvm_A	17	1	0,3599		0,9839		none	1	1
2	1ben_CD_2-HBD_BCD	1trz_ABCD	13	1	0,5557		0,9834		none	1	1
1	1did_A_1-DIG_A	1xla_A	17	1	0,3570		0,9834		none	1	1
1	1pnf_A_1-NAG-1-NDG_A	1png_A	13	1	0,4441		0,9833		none	1	1
1	1duc_A_1-DUD_A	1dun_A	12	1	0,4692		0,9823		none	1	1
1	1gmp_A_1-2GP_A	1gmq_A	12	1	0,4308		0,9816		none	1	1
1	1awb_A_1-IPD_A	2hhm_A	19	1	0,4094		0,9799		none	1	1
2	1ajo_A_1-PH2-1-SAN_A	1ajz_A	22	1	0,5151		0,9792		none	1	1
2	1mjl_AB_2-SAM_AB	1mjk_AB	32	1	1,1478		0,9788		rotamer	1	1
1	1bm7_A_1-FLF_A	1bmz_A	9	1	0,5365		0,9784		none	1	1
3	3gal_A_1-1GN_A	1bkz_AB	8	1	0,4852		0,9781		none	1	1
2	1igb_A_1-IPO_A	1amp_A	19	1	0,6915		0,9776		rotamer	1	1
2	1ddt_A_1-APU_A	1sgk_A	26	1	0,6854		0,9775		rotamer	1	1
2	1jdc_A_4-GLC_A	1jda_A	24	1	0,4877		0,9770		none	1	1
3	1swd_A_1-BTN_AD	1swa_ABCD	20	1	0,5165		0,9768		none	1	1
1	1kev_A_1-NDP_A	1ped_A	36	1	0,6332		0,9768		none	1	1
2	1epb_A_1-REA_A	1epa_AB	21	1	0,6416		0,9756		rotamer	1	1
1	1log_A_1-BMA-1-MAN-1-NAG_AB	1loe_A	13	1	0,6490		0,9746		none	1	1
1	1bso_A_1-BRC_A	2blg_A	17	1	0,7158		0,9745		none	1	1
3	2cht_A_1-TSA_AC	2chs_ABCDEFGHIJKL	13	1	0,5761		0,9726		none	1	1
3	1a9p_A_1-9DI_A	1a90_A	19	1	0,8366		0,9722		backbone	1	1
2	1kel_H_1-AAH_HL	1kem_HL	16	1	0,6290		0,9722		rotamer	1	1
1	1nft_A_1-NTA_A	1tfa_A	9	1	0,7784		0,9721		rotamer	1	1
2	1pgn_A_1-NBP-1-POP_A	2pgd_A	27	1	1,0848		0,9719		rotamer	1	1
2	1ubw_A_1-GPP_A	1ubv_A	17	1	0,7816		0,9700		loop	1	1
2	1ltt_D_1-BGC-1-GAL_D	1lts_ACDEFGH	10	1	0,6307		0,9697		rotamer	1	1
2	1aq7_B_1-34H-1-AG2-1-DIL-1-XPR_A	2ptn_A	27	1	0,7145		0,9693		loop	1	1
3	1cen_A_2-BGC_A	1ceo_A	15	1	1,2799		0,9687		loop	1	1
1	1br6_A_1-PT1_A	1rtc_A	15	1	0,9765		0,9674		rotamer	1	1
2	1ai2_A_1-ICA-1-NAP_A	3icd_A	34	1	0,7192		0,9672		loop	1	1
2	1mzm_A_1-PLM_A	1mzl_A	23	1	0,8714		0,9668		loop	1	1
3	1brp_A_1-RTL_A	1brq_A	27	0,93	1,0437		0,9658		loop	1	1
2	1fga_A_1-BME_A	2fgf_A	3	1	0,6379		0,9646		none	1	1
2	1fut_A_1-2GP_A	1fus_A	14	1	1,1221		0,9642		loop	1	1
2	1qpq_A_1-NTM_AB	1qpo_ABCDEF	12	1	0,9009		0,9633		rotamer	1	1
2	1nhk_R_1-CMP_R	2nck_LR	13	1	0,8294		0,9627		rotamer	1	1
2	1xva_A_1-SAM_AB	1bhj_AB	28	1	0,9473		0,9605		loop	1	1
2	1adl_A_1-ACD_A	1alb_A	21	1	0,7780		0,9601		backbone	1	1
2	1fkl_A_1-RAP_A	1fkk_A	14	1	0,8240		0,9597		loop	1	1
1	1tal_A_1-TAM_A	2alp_A	9	1	1,0782		0,9585		rotamer	1	1
2	1a26_A_1-CNA_A	2paw_A	13	1	0,7868		0,9562		loop	1	1
2	1jef_A_3-NAG_A	135l_A	13	1	1,0378		0,9550		loop	1	1
2	1ses_A_1-AHX_A	1sry_AB	23	1	0,9289		0,9515		loop	1	1
3	1lca_A_1-CB3-1-UMP_A	4tms_A	24	1	1,1880		0,9491		loop	1	1
2	1ra1_A_1-NAP_A	5dfr_A	19	1	0,8727		0,9482		rotamer	1	1
1	1rca_A_2-CGP_A	1aqp_A	15	1	1,3234		0,9473		rotamer	1	1
1	2enb_A_1-THP_A	1ena_A	15	1	1,1981		0,9470		loop	1	1
1	1icm_A_1-MYR_A	1ifb_A	18	1	0,8810		0,9468		none	1	1
3	1byb_A_4-GLC_A	1bya_A	37	0,97	2,0620		0,9461		loop	1	1
1	1arm_A_1-TRS_A	1yme_A	11	1	1,7848		0,9460		rotamer	1	1
2	1alw_A_1-ISA_A	1alv_AB	12	1	1,1983		0,9432		loop	1	1
1	1byq_A_1-ADP_A	1yer_A	18	1	1,5522		0,9426		loop	1	1
3	1hex_A_1-NAD_A	1xaa_A	20	1	1,3025		0,9421		backbone	1	1
1	1bxq_A_1-PP8_A	3app_A	27	1	1,3232		0,9407		backbone	1	1
2	1gd1_O_1-NAD_OR	2gd1_OPQR	30	1	1,1631		0,9407		loop	1	1
3	4fua_A_1-PGH_A	1fua_A	14	1	1,1548		0,9386		backbone	1	1
3	1aj7_H_1-NPE_HL	2rcs_HL	16	0,94	1,5830		0,9339		loop	1	1
3	2bgu_A_1-UDP_A	2bgt_A	23	1	1,2770		0,9257		backbone	1	1
3	1mpj_A_1-IPH_AB	3ins_ABCD	10	1	1,3569		0,9074		rotamer	1	1
3	1hnl_A_1-GSH_A	1lz4_A	10	0,8	1,9207		0,8863		loop	1	2
3	7tim_A_1-PGH_A	1ypi_AB	17	0,88	2,2689		0,8759		loop	1	2
3	1hii_B_1-C20_AB	1hsi_AB	32	1	2,2116		0,8722		backbone	1	2
3	1ebg_A_1-PAH_A	1ebh_AB	20	0.9	2,2267		0,8562		loop	1	2
3	1vrt_A_1-NVP_A	1rtj_AB	18	0,94	2,2539		0,8523		backbone	1	2
3	1oor_A_1-GDU_A	1fgx_AB	28	0,86	2,3605		0,8486		loop	1	2
3	5cna_A_1-MMA_A	1enq_ABCD	11	0,91	1,9822		0,8234		backbone	1	2
3	1anf_A_2-GLC_A	1omp_A	17	1/1	3,4017	0,9531	0,6499	0,9460	domain	2	3
3	1ake_A_1-AP5_A	4ake_AB	45	1/1	6,1382	1,5312	0,4395	0,9070	domain	3	3

Tabelle 12: EpitopeMatch, qualitative Untersuchung.

schluss der Domänenbewegungen (Abb. 47b, Abb. 48b) in den Bereichen der Klassen K_1^{EM} und K_2^{EM}.

Die Frage "How Different are Structurally Flexible and Rigid Binding Sites?" aus der Überschrift von Gunasekaran and Nussinov [57] lässt sich anhand der oben durchgeführten Analyse des Holo-Apo-Datensatzes wie folgt beantworten: Die Flexibilität der Bindetaschen ist durch die Bewegungen der Residuenreste, der kleineren Loops und die Domänenbewegungen gewährleistet. Allerdings ist der Anteil der für die Flexibilität der Bindetaschen verantwortlichen Residuen minimal, sodass die ausgewählten Bindetaschen ihre rigide Konformation größtenteils beibehalten und der Datensatz nur bedingt für die Untersuchung der Flexibilität geeignet ist. Dieselbe Frage wird im nächsten Abschnitt erneut aufgegriffen. Allerdings ist der Datensatz für die Untersuchung der Flexibilität bzw. der Rigidität das einzige Epitop-Paar 1ANF.A/1OMP.A aus der Klasse K_3^{EM} mit den deutlichen Unterschieden der Holo-Apo-Konformationen und die gesamte PDB.

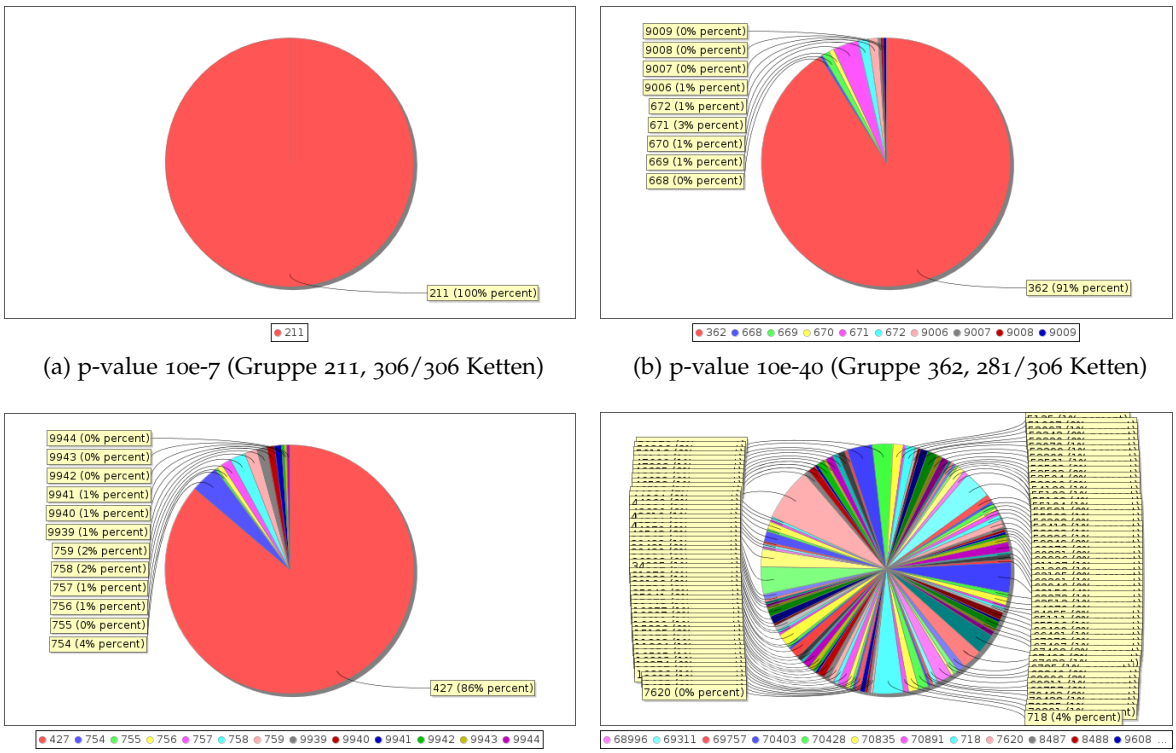
2.3.1.4 *Quantitativen Vergleiche - Epitop vs. PDB*

Als ein komparativer Test in [77] wurde der Vergleich mit dem Algorithmus ASSAM [12] durchgeführt. Im Unterschied zur damaligen Implementierung von EPITOPEMATCH basiert die aktuelle Version nicht auf einem graphentheoretischen, sondern auf einem kombinatorischen Ansatz. Darüber hinaus ist die neue Version um die vielfältigen Analysewerkzeuge erweitert. Die aktuelle und einzig zugängliche Version (online) des ASSAM [119] erlaubt neuerdings keine Suche nach Epitopen mit mehr als 12 Aminosäuren. Diese Gründe schließen das Wiederholen des komparativen Tests mit den beiden neuen Softwareversionen aus. Nichts desto trotz wird an dieser Stelle die Suche nach dem 1ANF.A-Epitop (Abb. 47) in der gesamten PDB erneut aufgegriffen um die Sensitivität, die Spezifität und die Analysemöglichkeiten von EPITOPEMATCH zu demonstrieren.

Der damalige (03.02.2009) Stand der PDB mit 55534 Strukturen ist heute (01.09.2014) auf 102863 Strukturen gewachsen. Jede einzelne der insgesamt 271962 Proteinketten aller 102863 Strukturen ist nach dem 17 Aminosäuren großen Epitop im Modus ALLATOMS/AHMW-NC7 durchsucht worden. Die Berechnung erfolgte auf 11 CPU-Kernen zweier Xeon X5650 innerhalb von 2h28m, sodass die Gesamtlaufzeit auf einem CPU-Kern ca. 27h in Anspruch nehmen würde. Gegenüber der Laufzeit der alten Version von 267h und angesichts fast der Verdopplung der PDB-Größe kann hier von einer ca. 20-fachen Performancesteigerung ausgegangen werden, bei einer deutlich umfangreicheren Ähnlichkeitsanalyse pro Match.

Mithilfe von EPITOPEMATCH werden die folgenden Fragestellungen untersucht: Wie variabel sind die Epitope der Maltodextrin bindenden Proteine im Rahmen ihrer Gruppeneinteilung innerhalb der Non-Redundant PDB chain set (NRPDB)? Existiert funktionale Verwandtschaft zu den nicht redundanten Strukturen innerhalb der PDB?

Die Bildung der NRPDB-Redundanzsets der Proteinketten basiert auf dem Sequenzalignment mit BLAST [7], synchronisiert mit dem Strukturalignment mit VAST [108]. Die Clusterung erfolgt nach BLAST p-value 10e-7, 10e-40, 10e-80, und 100% Sequenzidentität, woraus vier Redundanzlevels resultieren. Abb. 49 demonstriert die Verteilung der Kette A der Struktur 1ANF über vier Redundanzgruppen 211, 362, 427 und 718. Von insgesamt 306 redundanten Ketten der 182 Strukturen (Abb. 49a, niedrigste Redundanzwahrscheinlichkeit) liegen 35 im Komplex mit der Alpha-D-Glukose (GLC) vor, die in Form von unverzweigten Di-, Tri-, Hepta- oder Octasacchariden an die homologen Proteinketten gebunden ist. Der Ligand der Holo-Struktur 1ANF.A ist ein GLC-Disaccharid, der an den Kern des GLC-Epitops bindet (Abb. 47). Die Holo-Struktur 1ANF.A und die Apo-Struktur 1OMP.A mit den identischen Aminosäuren liegen trotz einer großen Konformationsänderung (Abb. 24)



(a) p-value 10e-7 (Gruppe 211, 306/306 Ketten)

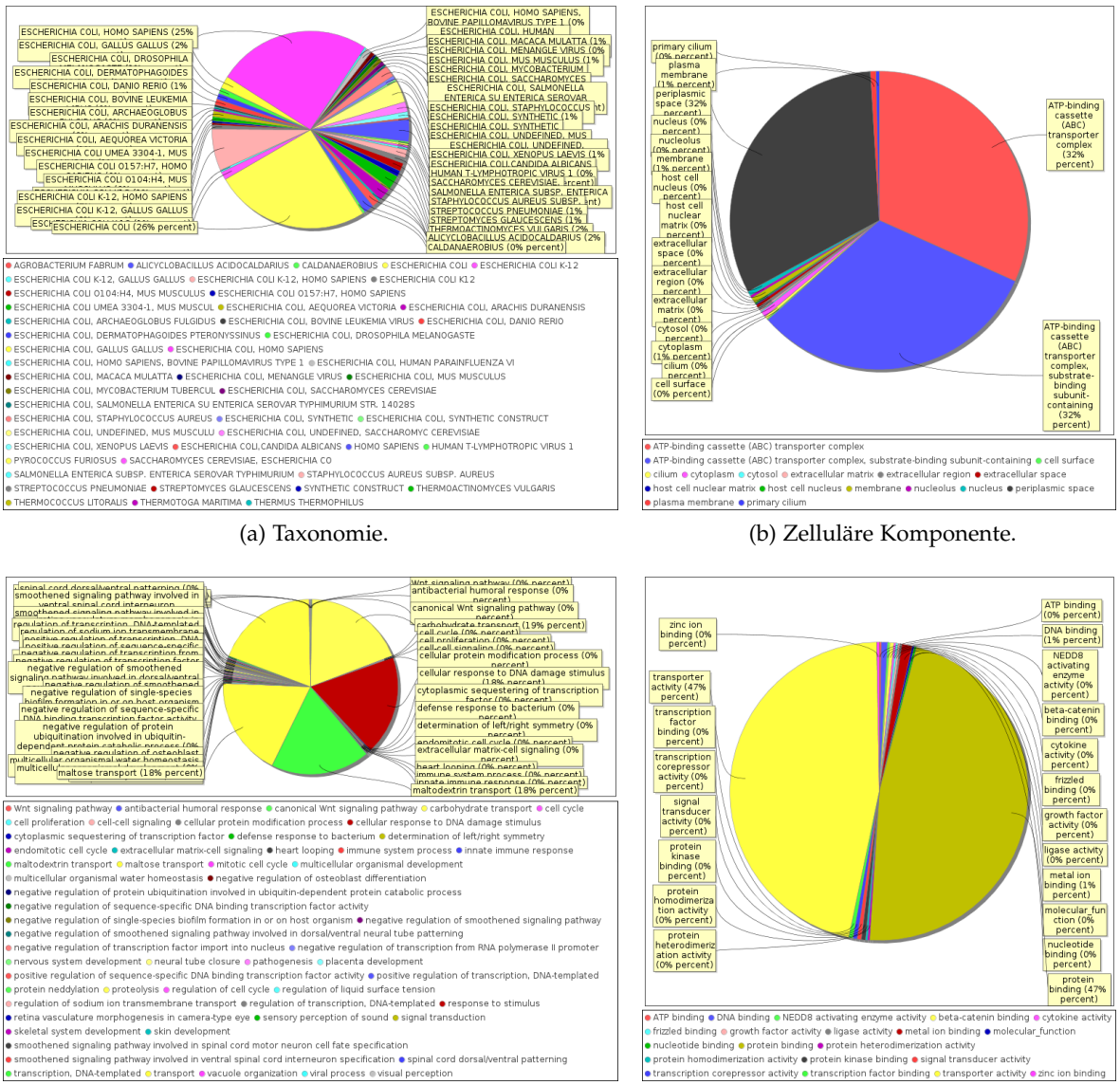
(b) p-value 10e-40 (Gruppe 362, 281/306 Ketten)

(c) p-value 10e-80 (Gruppe 427, 266/306 Ketten)

(d) nicht identisch (Gruppe 718, 12/306 Ketten)

Abbildung 49: 1ANF in der NRPDB. Stufenweise Einteilung der selben Struktur in die unterschiedlichen Redundanzgruppen: (a) die niedrigste Redundanzwahrscheinlichkeit bis (d) die höchste Redundanzwahrscheinlichkeit. Trotz der großen Unterschiede in der Konformation liegen die Holo-Struktur 1ANF.A in die Apo-Struktur 1OMP.A in der selben Gruppe (718) mit der höchsten Redundanzwahrscheinlichkeit (d). Die NRPDB unterscheidet nicht zwischen den Konformationsänderungen der redundanten Strukturen.

im selben Cluster mit der höchsten Redundanzwahrscheinlichkeit (Abb. 49d). Die NRPDB unterscheidet also nicht zwischen den Konformationsänderungen der redundanten Strukturen. Die homologen Strukturen werden innerhalb der jeweiligen Redundanzgruppe gemäß ihrem Rang geordnet. Der Rang der jeweiligen Struktur wird anhand der Anzahl der unbekannten und unvollständigen Residuen, der Strukturauflösung (X-ray), der Anzahl der Heterogene und der Residuen selbst ermittelt, sodass jede Gruppe eines Redundanzlevels durch eine Struktur mit dem besten Rang repräsentiert wird. Weder 1ANF.A noch 1OMP.A sind laut NRPDB repräsentativ. Allerdings liegen die beiden Strukturen innerhalb der größten Gruppen der mittleren Redundanzlevels (Abb. 49b, Abb. 49c) und innerhalb einer der vier größten Gruppen des höchsten Redundanzlevels (Abb. 49d), sodass sie durchaus als repräsentativ angesehen werden können. Der eigentliche Grund für die Wahl dieses Strukturpaares ist der kristallografischer Nachweis für ein großes Hinge-Twist bzw. Hinge-Bending [152], wobei das Induced-Fit das Hauptmerkmal dieser Untersuchung ist. Laut der Taxonomie des niedrigsten Redundanzlevels sind die meisten Ketten des bakteriellen Ursprungs (Escherichia coli, Abb. 50a). Sie befinden sich im periplasmatischen Raum Gram-negativer Bakterien und sind ein Teil des ABC-Transporter-Komplexes (GO, Abb. 50b). Die am häufigsten vorkommenden Biological Process (BP)s (GO, Abb. 50c) sind Maltodextrin-, Maltose- und Kohlenhydrat-Transport. Als Molecular Function (MF) (GO, Abb. 50d) stechen die Transportaktivität und die Proteinbindung heraus. 87 von 306 Ketten enthalten keine GO-Angaben,



(a) Taxonomie.

(b) Zelluläre Komponente.

(c) Biologischer Prozess.

(d) Molekulare Funktion.

Abbildung 50: Taxonomie & Gen-Ontologie der NRPDB-Gruppe 211. (a) Die meisten Ketten sind des bakteriellen Ursprungs (Escherichia coli). (b) Sie befinden sich im periplasmatischen Raum Gram-negativer Bakterien und sind ein Teil des ABC-Transporter-Komplexes. (c) Die am häufigsten genannten biologischen Prozesse sind Maltodextrin-, Maltose- und Kohlenhydrat-Transport. (d) Als molekulare Funktion stehen die Transportaktivität und die Proteinbindung im Vordergrund.

wobei 1ANF.A und 1OMP.A nicht darunter fallen. Mittels der Erkennung der Substrukturähnlichkeit können die GO-Angaben zumindest hypothetisch ergänzt werden.

Sensitivität

Zur Zeit der ersten Untersuchung [77] enthielt die PDB 70 Strukturen mit 96 homologen Ketten der Redundanzgruppe 211. EpitopeMatch erkannte die Holo-Epitope meist vollständig. Die Apo-Epitope waren jedoch nur teilweise erkannt, wobei die Ergebnisse relativ deckungsgleich ([77], Abb. 9) mit dem ebenfalls graphentheoretischen Ansatz von Assam [12] waren. Das Apo-Epitop auf 1OMP.A (Abb. 47b) wurde damals nicht erkannt. Die Cli-

quenerkennung reagiert empfindlich auf die großen Konformationsänderungen, was der Grund für den Umschwung auf den kombinatorischen Ansatz war.

Die 306 homologen Strukturen aus der NRPDB-Gruppe 211 (Abb. 49a) können als True Positive (TP) hinsichtlich der Kohlenhydrat-Bindung betrachtet werden. Das 1ANF.A-Epitop repräsentiert eine mögliche Holo-Konformation, an die das GLC-Disaccharid bindet, und somit die Funktion selbst. Wenn eine gefundene CS der NRPDB-Gruppe 211 dem 1ANF.A-Epitop entspricht, dann ist das Match TP. Andernfalls ist das Match False Negative (FN), da einer Struktur, die Kohlenhydrate binden kann, keine Kohlenhydrat-Bindungsfunktion in Form von 1ANF.A-Epitop zugeordnet werden konnte. In der Regel, wenn mindestens ein Punkt oder eine Gruppe der QS mindestens einem Punkt oder einer Gruppe der TS sowohl geometrisch als auch physiko-chemisch zugeordnet werden kann, liefert EpitopeMatch ein Match. Um die Ausgabe der Matches auf nur das Sinnvolle zu begrenzen, diente der QMCSS als ein Threshold, in dem die geometrische Ähnlichkeit (RMSD), physiko-chemische Ähnlichkeit (SSIM), die Größe des Epitops und die Ähnlichkeit der Substrukturen vereint sind (Gl. 45). Nur die Matches mit $QMCSS \geq 0.3$ sind berücksichtigt worden. Ob die Matches TP oder FN sind, kann mithilfe von Deskriptoren entschieden werden. Wenn die TP-CSs sich in den Zwielichtzonen der Deskriptorwerte befinden, dann hilft entweder die manuelle Prüfung oder die Kollisionserkennung. Die Kollisionserkennung reduziert zwar die manuelle Prüfung auf ein Minimum, sie erfordert jedoch die Existenz eines Liganden. Das Epitop muss also so wie im Fall von 1ANF.A ein Holo-Epitop sein. Bei der Kollisionserkennung wird jede QS anhand der Transformationsdaten der jeweiligen CS mit der TS überlagert und anschließend die Kollisionen der Atome des mittransformierten Liganden mit den Aminosäurenatomen der TS gezählt. Als eine Kollision gilt eine Distanz $d \leq 1.4\mathring{A}$ zwischen einem Atom des QS-Liganden und einem Atom der TS-Aminosäure. Holo- und Apo-Rotamere besitzen oft unterschiedliche Seitenkettenkonformationen. Aus diesem Grund werden 3 Kollisionen zugelassen. Für ein Epitop mit 17 Aminosäuren ist diese Anzahl von Kollisionen relativ gering. Abb. 51 zeigt die Verteilung der TP- und der FN-CSs bezüglich ihrer Grö-

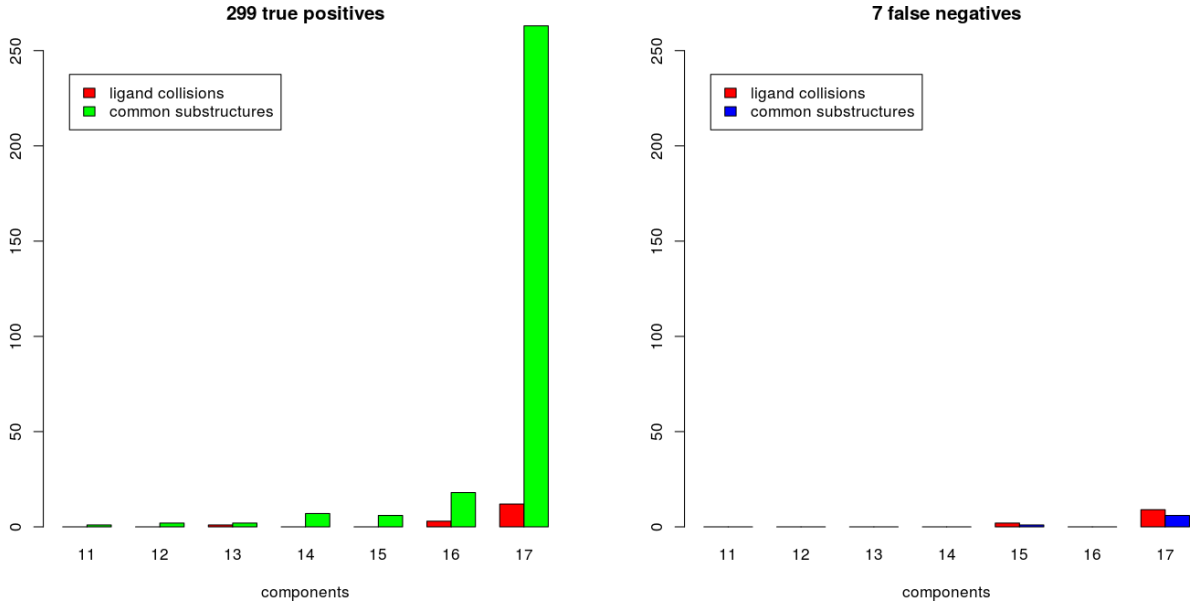
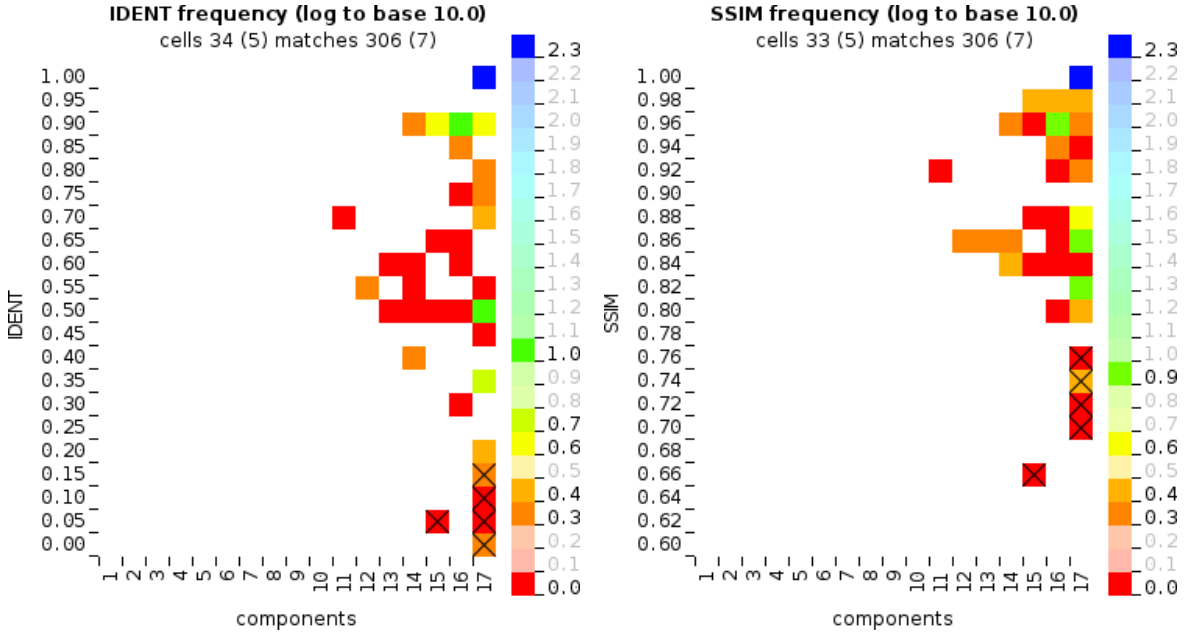


Abbildung 51: Sensitivität. Mittels der Transformationsdaten der gemeinsamen Substrukturen kann auch der Ligand der Holo-Struktur auf die Targetstruktur transformiert werden. Anhand der Anzahl der Kollisionen der Ligandatome mit den Atomen der Aminosäuren der Targetstruktur können die richtig-positiven CSs (minimale Anzahl von Kollisionen) identifiziert werden.

ße und der Anzahl der Kollisionen. Insgesamt 7 CSs sind FN. Auf den Strukturen 1EU8.A, 4G68.C, 4NDZ.A, 4QRZ.A, 4QSD.A, 4QSE.A und 4QSE.B konnte kein 1ANF.A-Epitop identifiziert werden. Diese CSs besitzen den größten Anteil von Kollisionen pro CS. Von insgesamt 299 TP-CSs sind 263 vollständig erkannt. Im Vergleich zu [77] (Abb. 9) mit etwa 35% der vollständig erkannten Holo-/Apo-Epitopen erkennt die neue Version also etwa 88% der Epitope vollständig. Eine Qualität, über die weder die alte EpitopeMatch-Version noch die alte Assam-Version verfügt hat. Die Sensitivität

$$sensitivity = \frac{TP}{TP + FN} = \frac{299}{299 + 7} \approx 0.977$$

für die Erkennung der Kohlenhydratbindungsfunktion der 306 homologen Strukturen anhand des 1ANF.A-Epitops liegt bei etwa 97.7%. Die Epitope der 7 FN-Strukturen sind im Vergleich zum 1ANF.A-Epitop mutiert und besitzen stark abweichende Konformationen, sodass CSs aus dem geometrischen Rauschen besser bewertet werden und die eigentlichen Epitope aufgrund der schlechten Bewertung aussortiert werden. Abb. 52a zeigt die Verteilung der Identität der CSs innerhalb der NRPDB-Gruppe 211. Die Frequenzskala ist logarithmisch. Die Werte der Skala sind Potenzen zur Basis 10, die für die jeweilige Anzahl der CSs in einem Identitätsbereich stehen. So sind 231 CSs ($10^{2.3} \leq 231 < 10^{2.4}$) vollständig erkannte Konformere des 1ANF.A-Epitops. Darunter fallen sowohl das in [77] nicht erkannte 1OMP.A- als auch die nur teilweise erkannten 1JW4.A-, 1JW5.A-, 2H25.A- und 2V93.A-Epitop, deren RMSD-Werte 3.41Å, 3.622Å, 3.569Å, 3.14Å und 2.826Å in der oder am Rande der Zwielichtzone (Abb. 10) liegen. Die 7 selektierten FN-CSs besitzen die Identität $IDENT < 0.2$. Die restlichen 68 TP-CSs sind entweder vollständig oder unvollständig



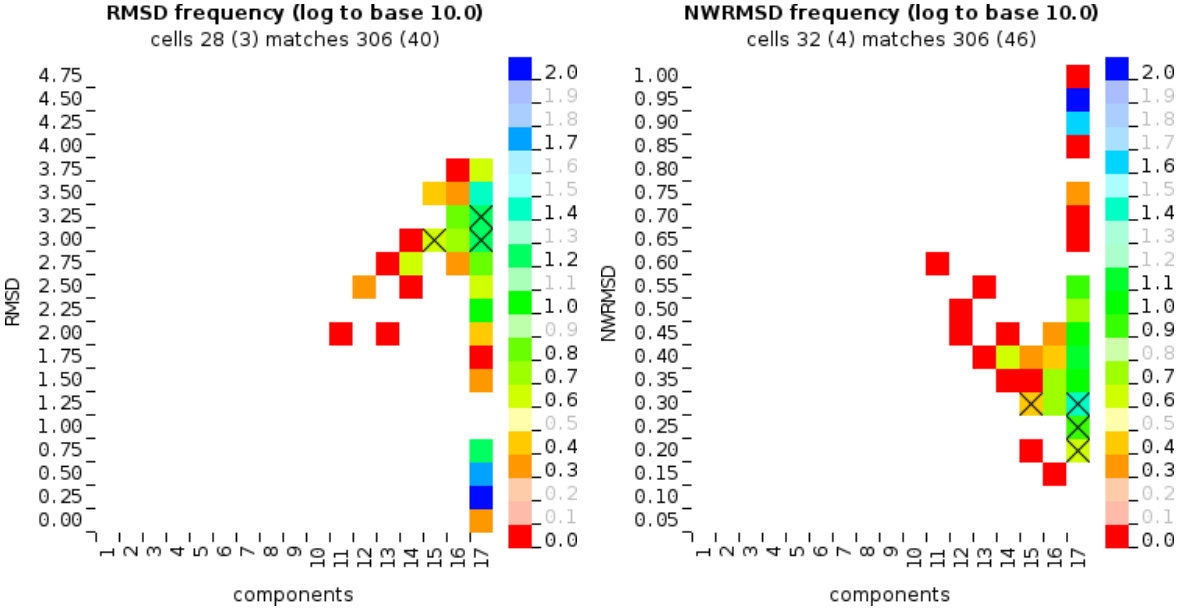
(a) IDENT (b) SSIM

Abbildung 52: IDENT & SSIM (NRPDB-Gruppe 211). Die logarithmische Farbskala kodiert die Frequenz der CSs im jeweiligen Deskriptorbereich. Die 7 falsch-negativen CSs (markierte Zellen) befinden sich in den untersten Bereichen der IDENT und der SSIM. 231 CSs ($10^{2.3} \leq 231 < 10^{2.4}$, blaue Zellen) sind vollständig erkannte Konformere des 1ANF.A-Epitops.

erkannt und besitzen mindestens eine Mutation gemessen an der Zusammensetzung des 1OMP.A-Epitops. Die Substitutionsähnlichkeit (Abb. 52b) setzt sich analog zu Gl. 63

$$ss_{ij}^{AH,MW,NC7} = \frac{3 - \frac{|diff_{ij}^{AH}|}{diff_{max}^{AH}} - \frac{|diff_{ij}^{MW}|}{diff_{max}^{MW}} - \frac{|diff_{ij}^{NC7}|}{diff_{max}^{NC7}}}{3} \tag{64}$$

aus AVEHYDROP, MOLWEIGHT und NETCHARGE zusammen, wobei die Substitutionswahrscheinlichkeit nach BLOSUM62 ausgelassen ist. Die 7 selektierten FN-CSs besitzen die Substitutionsähnlichkeit $SSIM < 0.8$. Die Trennung der TP-CSs von den FN-CSs mittels der beiden größenunabhängigen, physiko-chemischen Deskriptoren IDENT und SSIM ist durchaus möglich. Würde die Suche jedoch ohne der Berücksichtigung der Substitutionsähnlichkeit (Abb. 31) erfolgen, so müsste man sich nur auf die geometrische Ähnlichkeit stützen. Abb. 53a

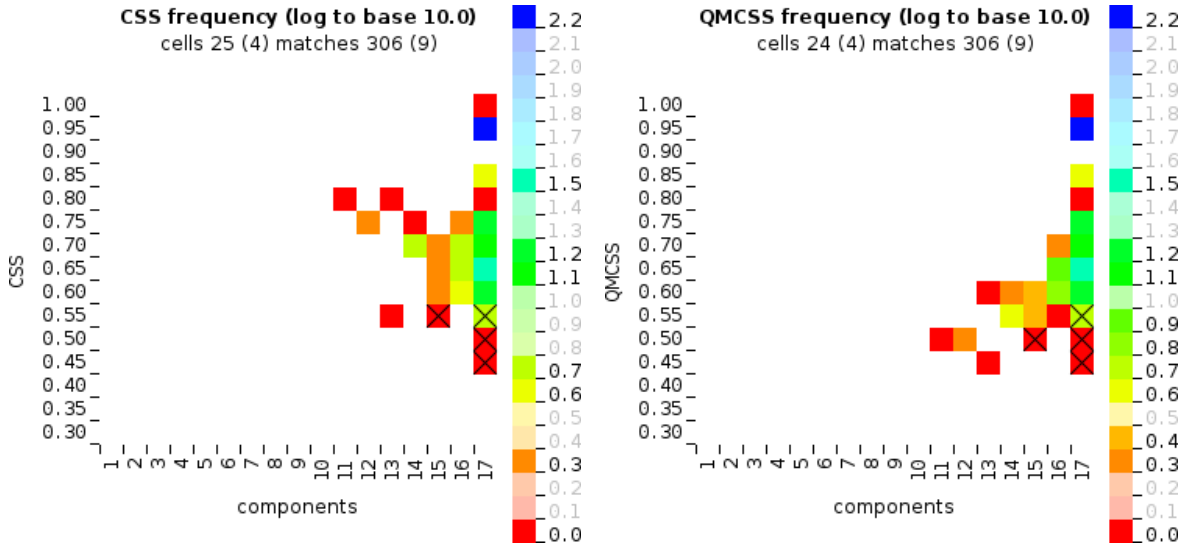


(a) RMSD (b) NWRMSD

Abbildung 53: RMSD & NWRMSD (NRPDB-Gruppe 211). Anhand der RMSD sind die 7 FN-CSs (markiert) von den Apo-CSs nicht zu unterscheiden. Die Verwendung der physiko-chemischen Gewichtung (NWRMSD) verschiebt sie in den unteren Bereich der Verteilung.

zeigt die Verteilung der RMSD-Werte. Anhand der Geometrie lassen sich 173 CSs in den Holo-Cluster mit $0.0Å \leq RMSD < 1.0Å$ und 133 CSs in den Apo-Cluster mit $1.5Å \leq RMSD < 4.0Å$ einordnen. Die 7 selektierten FN-CSs befinden sich mit $3.0Å \leq RMSD < 3.5Å$ "mitten" im Apo-Cluster und sind von den TP-CSs des Apo-Clusters praktisch ununterscheidbar. Der größenunabhängige, geometrische Deskriptor RMSD ist für die Trennung der TP-CSs von den FN-CSs ungeeignet und muss mit den physiko-chemischen Eigenschaften kombiniert werden. Die Kombination der Geometrie und der physiko-chemischen Eigenschaften (Abb. 53b) führt zu einer Verschiebung der 7 selektierten FN-CSs in den unteren Bereich der Verteilung. Sie befinden sich mit den 47 TP-CSs im Bereich $0.15 \leq NWRMSD < 0.35$ und sind somit von den TP-CSs, zwar deutlicher als im Fall von der RMSD, jedoch immer noch nicht eindeutig getrennt. Die alleinige Bewertung der Positionsabweichungen, kombiniert mit der Substitutionsähnlichkeit der korrespondierenden Residuen nach ihrer Überlagerung (NWRMSD), stößt beim Vergleich von Substrukturen mit dem großen Induced-Fit an ihre Grenzen. Jede CS setzt sich aus kleineren CSs zusammen. Zusätzliche Berücksichtigung der Substrukturinformationen der CSs erhöht die Messgenauigkeit der Ähnlichkeit.

In der Verteilung des größenunabhängigen CSS (Abb. 54a) befinden sich die 7 FN-CSs neben



(a) CSS (b) QMCSS

Abbildung 54: CSS & QMCSS (NRPDB-Gruppe 211). Erst das Einfließen der Bewertung der einzelnen Substrukturen in die Gesamtbewertung einer CS (CSS, Gl. 40) ermöglicht die eindeutige Trennung der TP- von den FN-CSs (markiert, im jeweiligen unteren Bereich).

3 TP-CSs im unteren Randbereich $0.45 \leq CSS < 0.6$. Auch hier sind die beiden Cluster: Holo $0.95 \leq CSS < 1.0$ und Apo $0.6 \leq CSS < 0.9$ gut zu erkennen. Die TP-CSs der Apo-Strukturen bestehen aus ähnlicheren Substrukturen als die FN-CSs. Anhand des Deskriptors CSS kann eine automatisierte Trennung der TP-CSs von den FN-CSs erfolgen. Die Anzahl der manuell zu überprüfenden Matches wird somit auf ein Minimum reduziert. Bringt man die Ähnlichkeit der CSs in den Zusammenhang mit ihrer Größe (Abb. 54b), so fallen alle CSs aus der Betrachtung heraus, die im Vergleich zu der gesuchten CS unangemessen klein sind.

Spezifität

Ausgehend von der Größe der NRPDB-Gruppe 211 mit 306 homologen Ketten (TPs) und der Gesamtzahl der Proteinketten in der PDB (271962) beläuft sich die Zahl der True Negative (TN)-Strukturen auf

$$TN = chains(PDB) - chains(NRPDB(211)) = 271962 - 306 = 271656$$

Ketten. Die False Positive (FP)-Ketten sind alle Ketten, die durch 1ANF.A-Epitop repräsentierte Kohlenhydratbindungsfunktion verfügen und einer anderen NRPDB-Gruppe als 211 zugeordnet sind. Dabei fungieren die FPs als Verbindungsglieder zwischen den unterschiedlichen Proteinfamilien mit der funktionalen Ähnlichkeit. Abb. 55 fasst die Verteilungen der CS-Deskriptoren für das 1ANF.A-Epitop in der PDB. Abb. 55a zeigt, dass die Zusammensetzung des 1ANF.A-Epitops in der vorliegenden Holo-Konformation sehr spezifisch ist. Die meisten TN-CSs befinden sich in den Zwielichtzonen der Deskriptoren $0.0 \leq IDENT < 0.2$ (Abb. 55a), $0.6 \leq SSIM < 0.8$ (Abb. 55b), $2.5Å \leq RMSD < 4.0Å$ (Abb. 55c), $0.15 \leq NWRMSD < 0.4$ (Abb. 55d), $0.4 \leq CSS < 0.6$ (Abb. 55e) und $0.4 \leq QMCSS < 0.6$ (Abb. 55f). Die Bereiche der TN-CSs überschneiden sich mit den entsprechenden Bereichen der 7 FN-CSs (Abb. 52, Abb. 53, Abb. 54) und markieren zugleich die Zonen mit dem höchsten physiko-chemischen und/oder geometrischen Rauschen. Die Ober- und die Untergrenzen der Zwielichtzonen sind lediglich Richtwerte, auf

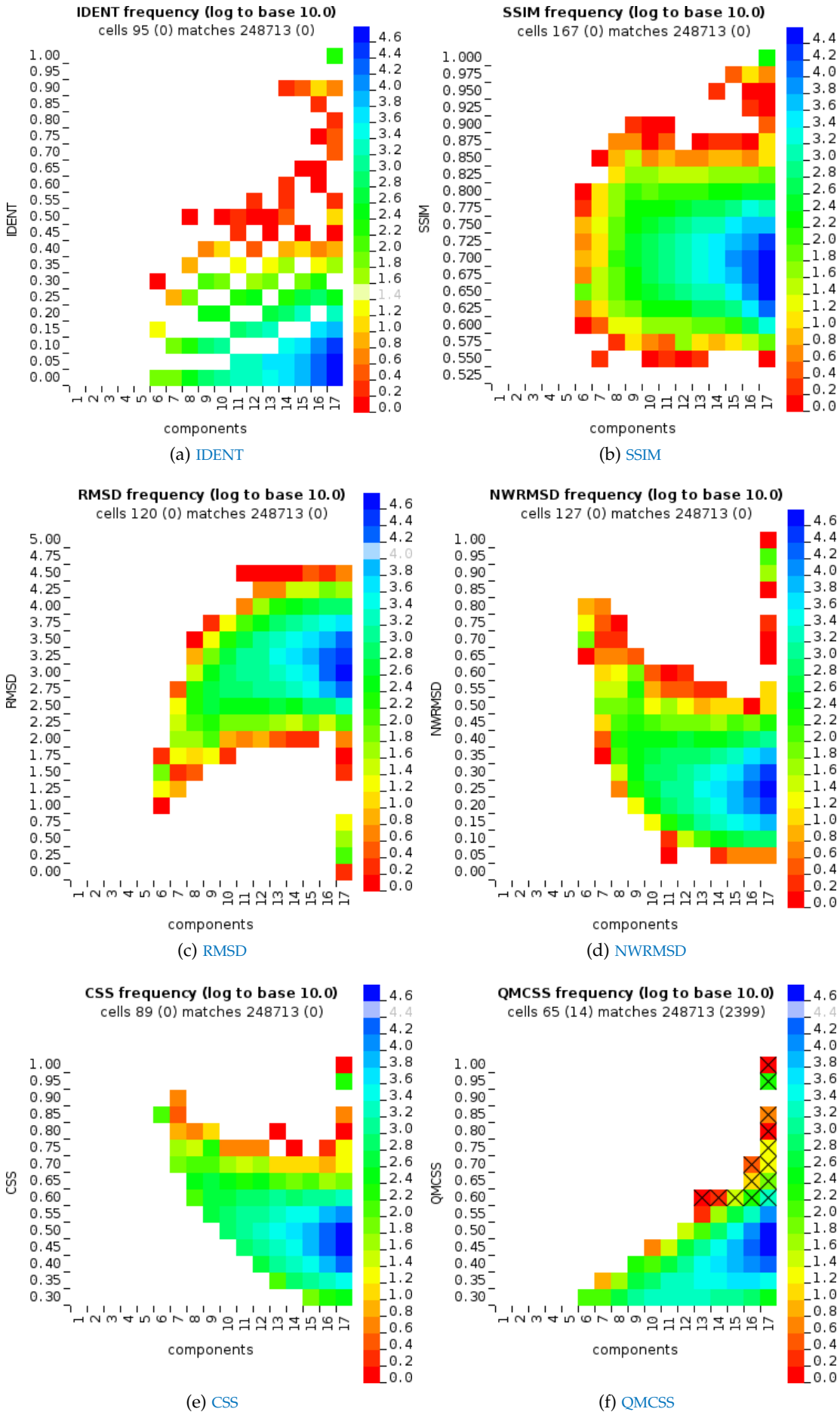
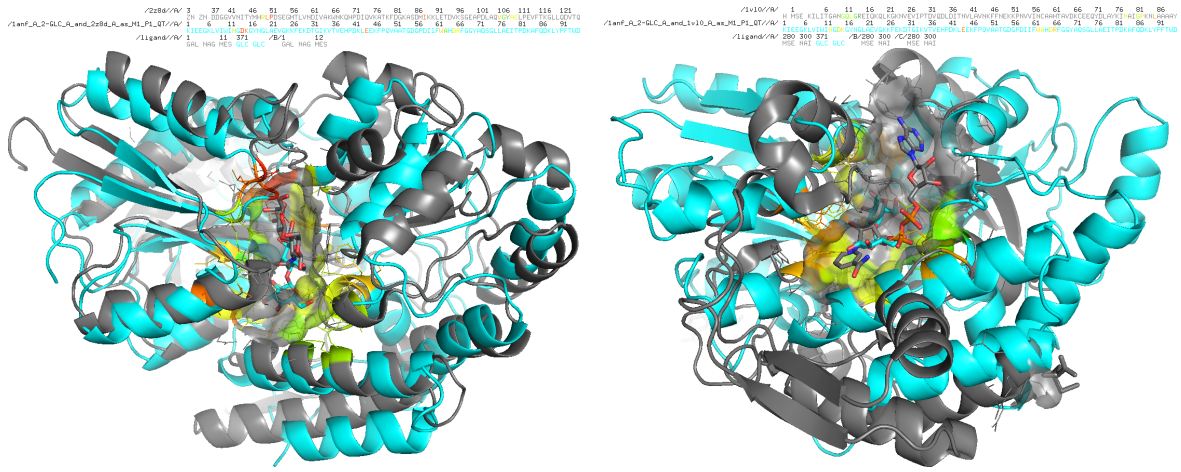


Abbildung 55: Deskriptorfrequenzen des 1ANF.A-Epitops in der PDB.

die man sich bei der Ergebnisfilterung stützen kann. Abb. 55f zeigt, dass in dem Bereich $0.6 \leq QMCSS < 1.0$ neben den 299 TP-CSs 2100 weitere CSs enthalten sind. Das Rauschen setzt ab etwa $QMCSS < 0.7$ ein. Diese Überschneidung der TP- und der TN-CSs ist die Folge des starken Induced-Fit (Abb. 47) des gesuchten Epitops, der sich auf zwei Domänen verteilt, die im Prozess der Wechselwirkung an der Bindungsstelle relativ weit auseinander driften. SCOP (v1.75) klassifiziert 1ANF.A als eine einzige Domäne. CATH (v3.5.0) jedoch unterteilt 1ANF.A in 2 Domänen. Im Gegensatz zu dem starren Deskriptor QMCSS ist der DMCSS flexibel. Während der $QMCSS = 0.673$ des 1ANF.A/1OMP.A-Epitops sich im verrauschten Bereich befindet, würde der flexible $DMCSS = 0.946$ (Tab. 12, K_3^{EM}) die Apo-CS aus dem Rauschen hervorheben und ähnlich spezifisch wie die Holo-CSs bewerten. Das gilt auch für die restlichen Apo-Epitope, die sich analog zu 1ANF.A/1OMP.A in den TP-Bereich der QMCSS-Verteilung verschieben würden. Die Implementierung der Ausgabe der Domain-Scores ist für die kommende Version vorgesehen. Nach der Durchführung der Kollisionserkennung auf 2100 CSs aus dem Bereich $0.6 \leq QMCSS < 1.0$ sind 12 FP-CSs erkannt worden. Sodass von einer Spezifität

$$specificity = \frac{TN}{TN + FP} = \frac{271656}{271656 + 12} \approx 1.0$$

gesprochen werden kann. Strukturen 3DMo.A und 4LOG.A (NRPDB-Gruppen 62 und 99, mit $QMCSS = 0.979$ und $QMCSS = 0.712$, beide mit $SSIM = 1.0$) verfügen über identische Epitope zu 1ANF.A. Beide Strukturen sind Fusionen aus dem maltosebindenden Protein (MBP) und einem Rezeptor für die aktivierte C-Kinase 1 (RACK1) im Fall von 3DMo.A [171] bzw. einem fotorezeptorspezifischen nuklearen Rezeptor (NR2E3/PNR) im Fall von 4LOG.A [170]. Die Fusion mit dem MBP wirkt stabilisierend auf RACK1 bzw. NR2E3 und erleichtert die Kristallbildung. Dennoch müsste jede der beiden Strukturen außer in der NRPDB-Gruppe 62 bzw. 99 auch in der NRPDB-Gruppe 211 enthalten sein, da die MBP-Struktur in beiden Fällen mit enthalten ist. Die eineindeutige Gruppierung der Strukturen bzw. die fehlende Domäneneinteilung ist ein weiterer Nachteil der NRPDB. Die weiteren 5 Strukturen mit ähn-



(a) 2Z8D.A, entfernt homologe Strukturen. (b) 1VLo.A, nicht homologe Strukturen.

Abbildung 56: Grenze der faltungsmusterabhängigen Ähnlichkeit.

lichen Epitopen zu 1ANF.A sind 2HPG.B ($QMCSS = 0.605$, $NRPDB(10e-7) = 3658$) und 4I1D.A-D ($0.641 \leq QMCSS \leq 0.663$, $NRPDB(10e-7) = 3376$), die als ABC-Transporter bekannt sind. Bei 2 weiteren Strukturen 2Z8D.A (Abb. 56a) und 2Z8F.B ($QMCSS = 0.631$ und $QMCSS = 0.627$, beide $NRPDB(10e-7) = 11354$) handelt es sich um zuckerbinden-

den Proteine, die statt der Maltose, Galacto-N-Biose transportieren. Laut CATH verfügen alle 9 bisher genannten Strukturen über ein Domäne der Klasse "Alpha Beta", mit der Architektur eines "3-Layer(aba) Sandwich", mit der Topologie eines "D-Maltodextrin-Binding Protein" und gehören der Superfamilie "Periplasmic binding protein-like II" an. Somit ist die Ähnlichkeit ihrer Epitope zum gesuchten 1ANF.A-Epitop (positive) als FP im Sinne ihrer Einteilung in die unterschiedlichen NRPDB-Gruppen (false) zu betrachten. Bei den 3 letzten Strukturen 1VL0.A-C (Abb. 56b, $QMCSS = 0.652$, $NRPDB(10e-7) = 3022$) handelt es sich um eine Domäne der gleichen Klasse und Architektur. Die Topologie ist allerdings "Rossmann fold" und die Superfamilie "NAD(P)-binding Rossmann-like Domain", sodass die dem 1ANF.A-Epitop ähnliche Konformation an der Oberfläche des 1VL0.A sowohl als FP als auch als TN bewertet werden kann. Die kollisionsfreie, ähnliche Position der Maltose im 2Z8D.A-GAL-NAG-Epitop (Abb. 56a), ähnliche Liganden (Disaccharide GLC-GLC und GAL-NAG), die ähnliche Überlagerung der beiden vollständigen Strukturen 1ANF.A und 2Z8D.A anhand der Transformationsdaten der CS und die gleiche Domänen-Klassifizierung nach CATH zeugen von einer funktionalen Verwandtschaft der NRPDB-Gruppen 211 (306 Ketten) und 11354 (6 Ketten). Im Fall des 1VL0.A-NAI-Epitops (Abb. 56b) scheint die Ähnlichkeit der CS eher zufallsbedingt zu sein, zumal lediglich 3 der insgesamt 27 Strukturen der NRPDB-Gruppe 3022 über diese Ähnlichkeit verfügen. Die Beispiele in der Abb. 56 veranschaulichen die Erkennungsgrenze der faltungsmusterabhängigen Ähnlichkeit. Trotz der Verfügbarkeit einiger wenigen FP-CSs liegt die Spezifität für die Erkennung des 1ANF.A-Epitops auf dem Hintergrund der gesamten PDB bei etwa 99.996%.

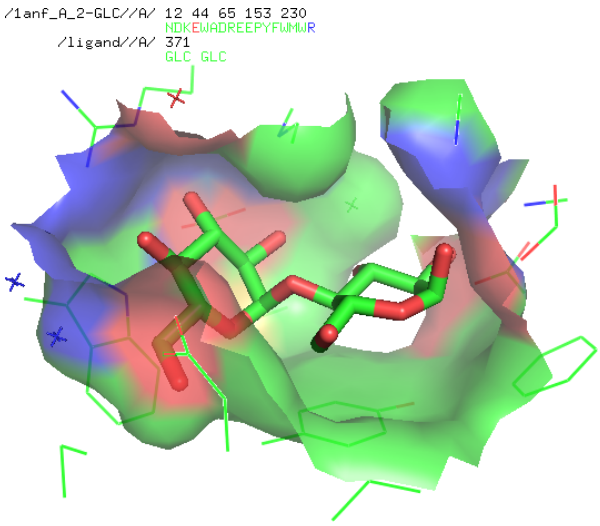
2.3.1.5 *Hotspots*

Der erste Schritt in die Unabhängigkeit von dem Faltungsmuster ist die Verwendung von Hotspots. Die bisher vorgestellten AllAtoms- und Backbone-Templates sind faltungsmustergebunden, weil sie Korrespondenzen zwischen den Rückgratatomen enthalten. Entfernt man die Rückgratatome N, Cα, C' und O aus dem AllAtoms-Template (Abb. 5), so gelangt man zu einem SideChains-Template, aus dem das Glycin aufgrund der fehlenden Seitenkette herausfallen würde. Die Seitenketten lassen sich ferner in Atomgruppen splitten. Jede Atomgruppe verfügt über die, für die molekulare Interaktion wichtigen, physikochemischen Eigenschaften AC, DO, PI, AL und DA [149]. Tab. 13 enthält die entsprechenden Atomgruppen, deren geometrische Zentren die Pseudozentren mit der jeweiligen Eigenschaft repräsentieren. Die zu einer Gruppe gehörenden Atome sind durch Kommata getrennt. Die einzelnen Atomgruppen sind durch Leerzeichen getrennt. Die Peptidbindung wird durch die Gruppen AC (O), DO (N) und PI (C') repräsentiert. Die darauf basierende Substitutionsmatrix lässt nur Korrespondenzen zwischen den Hotspots des gleichen Typs zu und ist somit spezifisch. Zerlegung der Seitenketten in Hotspots führt zu einer höheren Auflösung des jeweiligen Epitops. Abb. 57a zeigt die Atome der Hotspots und ihre Oberfläche aus der 5.0$\mathring{A}$-Umgebung der Maltose aus 1ANF.A. Das 17 Aminosäuren große 1ANF.A-Epitop lässt sich durch insgesamt 30 Seitenketten-Hotspots (Abb. 57b, oliv) beschreiben, wobei die Hotspots der Peptidbindungen bzw. des Rückgrats nicht berücksichtigt werden. Die Hotspots einer Struktur beteiligen sich an den internen und/oder externen Interaktionen. Unter der Annahme, dass die meisten internen Interaktionen im inneren der Struktur und die meisten externen Interaktionen an der SAS der Struktur stattfinden, können die an der SAS unbeteiligten Hotspots der jeweiligen TS ausgeschlossen werden. EpitopeMatch implementiert eine Schnittstelle zum Maximal Speed Molecular Surface (MSMS)-Programm [147], dass für ein Set von Kugeln, unter der Angabe des Probe-Radius ($r = 1.4\mathring{A}$) und der Vertexdichte ($d = 3/\mathring{A}^2$), analytische Modelle der SAS und der Solvent Excluded Sur-

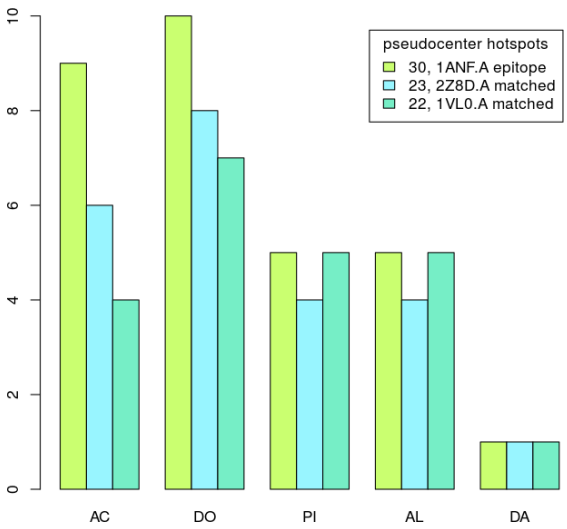
AS	AC	DO	PI	AL	DA
ALA	O	N	C′	Cβ	
ARG	O	N Nε Nη1 Nη2	C′	Cβ,Cγ,Cδ	
ASN	O Oδ1	N Nδ2	C′		
ASP	O Oδ1 Oδ2	N	C′		
CYS	O	N	C′	Cβ,Sγ	
GLN	O Oε1	N Nε2	C′		
GLU	O Oε1 Oε2	N	C′		
GLY	O	N	C′		
HIS	O	N	C′ Cγ,Nδ1,Cδ2,Cε1,Nε2		Nδ1 Nε2
ILE	O	N	C′	Cβ,Cγ1,Cγ2,Cδ1	
LEU	O	N	C′	Cβ,Cγ,Cδ1,Cδ2	
LYS	O	N Nζ	C′	Cβ,Cγ,Cδ,Cε	
MET	O	N	C′	Cβ,Cγ,Cδ,Cε	
PHE	O	N	C′ Cγ,Cδ1,Cδ2,Cε1,Cε2,Cζ		
PRO	O	N	C′	Cβ,Cγ,Cδ	
SER	O	N	C′		Oγ
THR	O	N	C′	Cγ2	Oγ1
TRP	O	N Nε1	C′ Cγ,Cδ1,Cδ2,Nε1,Cε2,Cε3,Cζ2,Cζ3,Cη2		
TYR	O	N	C′ Cγ,Cδ1,Cδ2,Cε1,Cε2,Cζ		Oη
VAL	O	N	C′	Cβ,Cγ1,Cγ2	

Tabelle 13: Pseudozentren der Hotspots nach [149] mit den Typen Wasserstoffbrückenbindung-Akzeptor (AC), Wasserstoffbrückenbindung-Donor (DO), aromatischer Pi-Kontakt (PI), hydrophob aliphatisch (AL) und Donor/Akzeptor gemischt (DA).

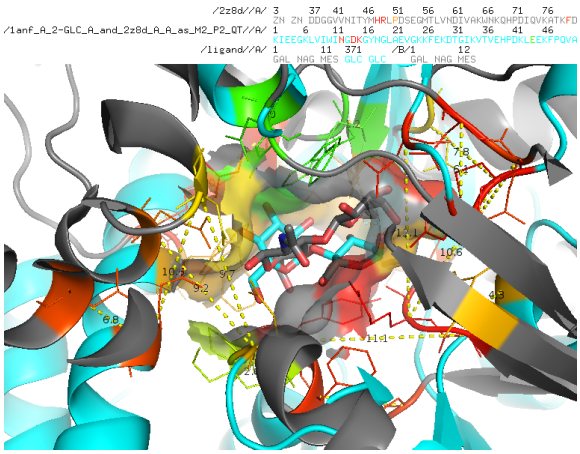
face (SES) erstellt. MSMS gibt pro Atom die SAS- und die SES-Fläche in $Å^2$ an. Alle Atome mit der SAS- bzw. SES-Fläche $A = 0Å^2$ werden ausgeschlossen. Die Koordinaten der verbleibenden Atome mit der SAS- bzw. SES-Fläche $A > 0Å^2$ werden anhand des HOTSPOTS-Templates (Tab. 13) in Form von Hotspots zusammengefasst, wobei jeder Hotspot durch das geometrische Zentrum der zusammengehörenden Hotspotatome der jeweiligen Aminosäure repräsentiert wird. Abb. 57c demonstriert die Hotspot-CS des 1ANF.A-Epitops und der durch Hotspots repräsentierten Oberfläche der 2Z8D.A-TS. Die resultierende Überlagerung der beiden Holo-Strukturen anhand der Transformationsdaten der gefundenen CS ($HOTSPOTS = 23$, $RMSD = 4.148Å$, $MCSS = 0.398$) offenbart eine Gemeinsamkeit an den Oberflächen der beiden Strukturen und bringt die beiden Disaccharide in die relativ ähnliche Position innerhalb der CS. Die gestrichelten gelben Linien verdeutlichen die Entfernungen zwischen den Cα-Atomen der Aminosäuren mit den korrespondierenden Hotspots. Im Fall der 1VLo.A-TS (Abb. 57d) wird eine CS mit $HOTSPOTS = 22$, $RMSD = 3.837Å$, $MCSS = 0.391$ erkannt. Die Oberflächendarstellung bezieht sich auf die Kontakte der jeweiligen TS und des transformierten Disaccharids der QS. Beide CSs weisen eine relativ hohe RMSD auf, die jedoch aufgrund der Mittelung der Atomkoordinaten der Hotspots und des ggf. mehrfachen Vorkommens der Hotspots pro Seitenkette durchaus legitim ist. Die Histogramme der CSs (Abb. 57b) zeigen, dass die 1VLo.A-CS mehr gemeinsame PI- und AL-



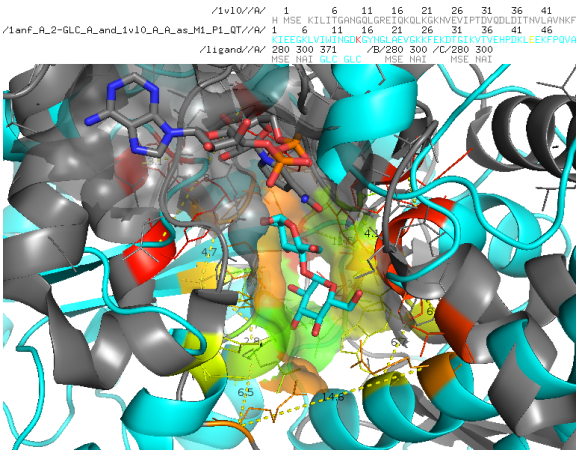
(a) 30 Hotspots des 1ANF.A-Epitops.



(b) Gematchten 2Z8D.A- und 1VLo.A-Hotspots.



(c) Gematchten 2Z8D.A-Hotspots.



(d) Gematchten 1VLo.A-Hotspots.

Abbildung 57: Pseudozentren-Hotspots. Die Zusammenfassung der Seitenketten als Hotspots ermöglicht die Trennung der CSs von dem Faltungsmuster.

Hotspots (10) enthält als die 2Z8D.A-CS (8) und ist somit insgesamt hydrophober. Auf der anderen Hand enthält die 2Z8D.A-CS mehr AC- und DO-Hotspots (14) als die 1VLo.A-CS (11) und ist somit hydrophiler. Weder die 2Z8D.A-CS noch die 1VLo.A-CS sind vollständig. Beide existieren jedoch und bestärken somit die faltungsmusterabhängige Ähnlichkeit (Abb. 56) zu dem 1ANF.A-Epitop. Die 2Z8D.A-CS kann mit einem Hotspot mehr und einer ausgeglicheneren Verteilung der korrespondierenden Hotspots als ähnlicher betrachtet werden. Dieser Ansatz ist analog zu SiteEngines [160, 161], das auf einem GH-Algorithmus basiert. Die Umstellung auf das Matchen von Hotspots bedeutet für EpitopeMatch lediglich die Definition des Hotspots-Templates (Tab. 13). Der Algorithmuskern und die Algorithmusparameter (Abs. 2.2) bleiben unverändert.

2.3.1.6 *Vertexnormalen*

EpitopeMatch beschränkt sich nicht auf die Koordinaten der Atome. Der Schritt in die vollständige Unabhängigkeit von den Faltungsmustern ist die Verwendung der Vertexkoordinaten der molekularen Oberflächen. MSMS trianguliert das analytische Modell der SES, die im Fall von 1ANF.A mit dem Probe-Radius $r = 1.4\mathring{A}$ und der Vertexdichte $D = 3/\mathring{A}^2$

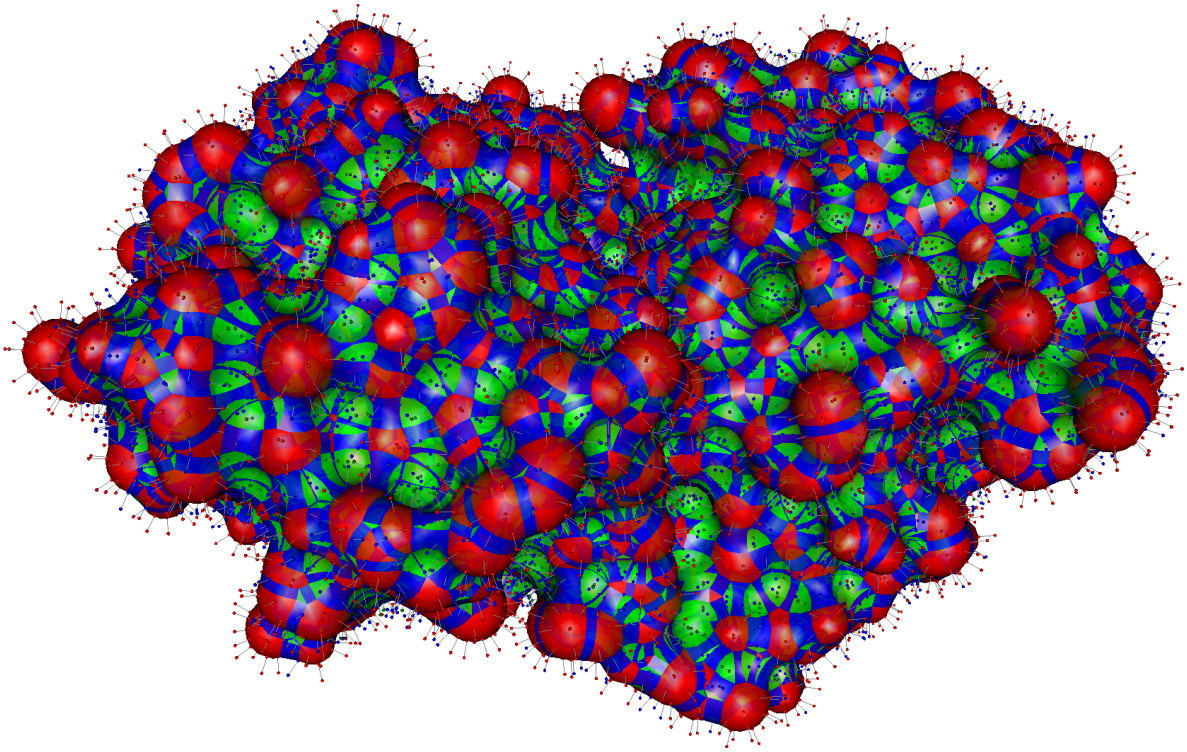


Abbildung 58: Triangulierte Oberfläche der 1ANF.A mit den Kontaktflächen CF (rot), sphärischen Flächen SF (grün) und torischen Flächen TF (blau). Die tatsächliche Vertexdichte $D = 3/\mathring{A}^2$ der dargestellten Flächen ist auf $D \approx 0.534/\mathring{A}^2$ reduziert. Zu jedem Vertex (Kugeln unmittelbar an der Oberfläche) gehört eine Normale (über die graue Linie mit dem jeweiligen Vertex verbundenen Kugeln). Die Normalen zeigen die Außenseite der Oberfläche an.

$A^{SES} = 13942.147\mathring{A}^2$ beträgt. Die resultierende SES wird durch 37285 Vertices beschrieben ($A^{SES} \approx 37285/D$). Jeweils 3 Vertices bilden eine Fläche (Face), die entweder CF, SF oder TF (Abb. 58) ist. EpitopeMatch schließt die Vertex-Nachbarschaften von $\text{dist}(V_i, V_j) < 1.0\mathring{A}$ aus, verringert somit die Anzahl von repräsentativen Vertices pro Face und erreicht eine gleichmäßigere Verteilung der Vertices, deren Anzahl sich auf insgesamt 7444 reduziert (Abb. 58, nach den Farben der entsprechenden Flächen gefärbten Kugeln unmittelbar an der Oberfläche). Die resultierende Vertexdichte ist $D \approx 0.534/\mathring{A}^2$. Zu jedem Vertex gehört eine Normale, die aus den Normalen der angrenzenden Flächen (Dreiecke) gemittelt wird (Abb. 58, nach den Flächen gefärbte Kugeln, $1.0\mathring{A}$ von dem jeweiligen Oberflächenvertex entfernt und mit demselben über die graue Linie verbunden). Eine Vertexnormale ist also ein Koordinatenpaar, das die Position und Ausrichtung eines Oberflächenpunktes repräsentiert. Insgesamt 7444 Vertexnormalen beschreiben somit die SES-Topologie der 1410 Atome (von insgesamt 2860 Atomen der 1ANF.A) mit der SES $A > 0\mathring{A}^2$ (Abb. 59). Je größer die SES eines Atoms, desto mehr Vertexnormalen beschreiben ihn. Analog zum Cα-Cβ-Template, das die faltungsmusterabhängige Position einer Aminosäure und die Ausrichtung ihrer Seitenkette beschreibt, definiert ein Vertexnormalen-Template die faltungsmusterunabhängige Position und Ausrichtung eines Oberflächenhotspots auf der SES. Abb. 60a zeigt 130 Oberflächenhotspots aus der $3\mathring{A}$-Umgebung der Maltose. Die Oberfläche des Epitops beträgt $A = |V|/D = 130/0.534/\mathring{A}^2 = 243.446\mathring{A}^2$ und ist somit ca. 1.75% der Gesamtoberfläche. Die Geometrie des Epitops wird durch 130 Vertexnormalen repräsentiert. Die physikochemischen Eigenschaften der Oberflächenhotspots werden nach Tab. 13 zugeordnet, woraus sich 25 AC (rot), 30 DO (blau), 40 PI (cyan), 34 AL (grün) und 1 DA (magenta) ergeben. Die

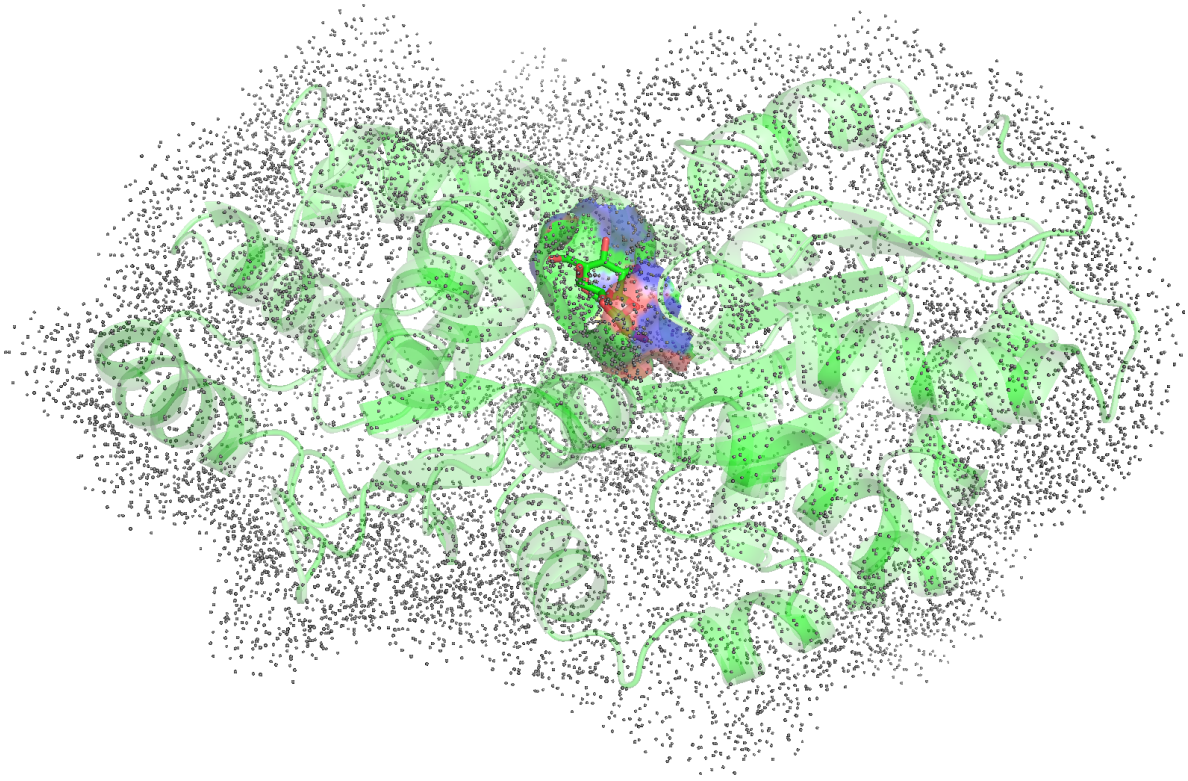
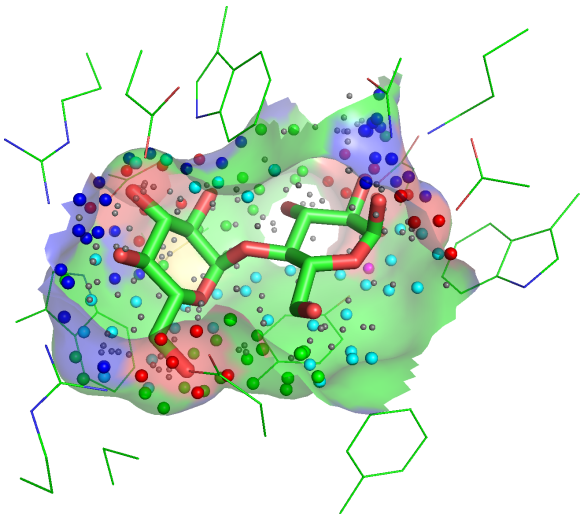
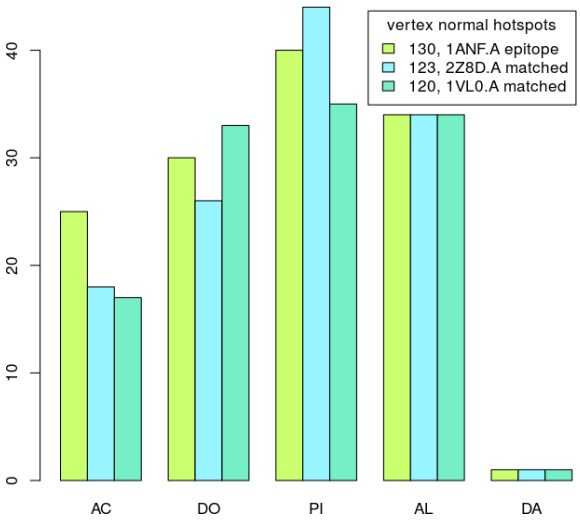


Abbildung 59: Auf die Vertexnormalen (Kugelpaare aus Abb. 58) reduzierte Oberfläche der 1ANF.A. Jede Vertexnormale repräsentiert die faltungsmusterunabhängige Position und Ausrichtung eines Oberflächenhotspots. Je größer ein Hotspot ist, desto mehr Vertexnormalen beschreiben ihn.

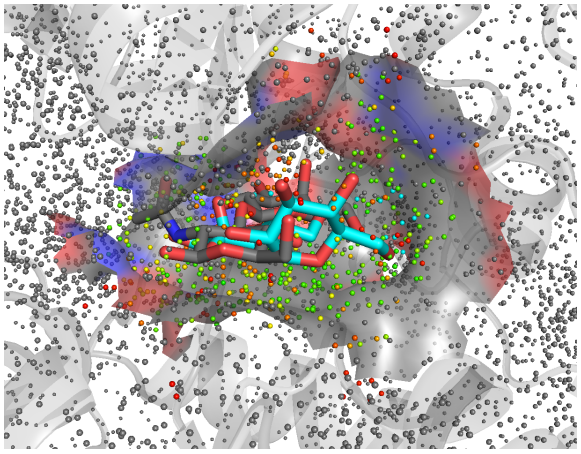
Verteilung der physiko-chemischen Anteile der Oberflächenhotspots (Abb. 60b, oliv) zeigt, dass die Anzahl der aromatischen und aliphatischen Kontakte (PI und AL) der Maltose und der 1ANF.A größer als die Anzahl der Wasserstoffbrückenbindungskontakte (AC , DO und DA) ist. Dieses Detail ist aus der Definition der Hotspots durch die geometrischen Zentren der Seitenkettenatome (Pseudozentren, Abb. 57b, oliv) nicht direkt ersichtlich. Die Repräsentation der Epitope anhand der Oberflächenhotspots bietet also neben einer vollständigen Trennung von dem Faltungsmuster eine deutlich höhere Auflösung des Informationsgehalts. EpitopeMatch erkennt auf der Oberfläche von 2Z8D.A eine MCS mit $HOTSPOTS = 123$, $RMSD = 2.898\mathring{A}$ und $MCSS = 0.703$ (Abb. 60c). 34.15% der Hotspots korrespondieren mit den identischen Aminosäuren des 1ANF.A-Epitops. Die Oberfläche von 1VLo.A enthält eine MCS mit $HOTSPOTS = 120$, $RMSD = 3.069\mathring{A}$, $MCSS = 0.644$ und 26.67% Hotspots mit den identischen Aminosäuren (Abb. 60d). Die Bewertung der beiden MCSs fällt deutlich besser aus als im Fall der Suche anhand der Pseudozentren-Hotspots. Die höhere Anzahl der erkannten PI-Hotspots auf 2Z8D.A (Abb. 60b, blau, 44) als die Anzahl der gesuchten PI-Hotspots (Abb. 60b, oliv, 40) erklärt sich durch die Besonderheit des Hotspots-Templates (Tab. 13), in dem der Nε1-Atom eines TRP sowohl die PI- als auch die DO-Eigenschaft besitzt. Das bedeutet, dass der Nε1-Atom des TRP-Rotamers an der 2Z8D.A-Oberfläche, relativ zu den restlichen Atomen des Epitops, weniger oder anders exponiert ist, als an der Oberfläche der 1ANF.A. Der Trend für den Mangel an AC- und DO-Hotspots im Fall der Suche nach den Pseudozentren-Hotspots (Abb. 57b) geht bei der Suche nach den Vertexnormalen-Hotspots (Abb. 60b) in die umgekehrte Richtung. Die Bewertung der beiden MCSs fallen mit 70.3% (2Z8D.A) und 64.4% (1VLo.A) deutlich höher aus, als bei der Suche mittels Pseudozentren-Hotspots (39.8% und 39.1%). Demnach ist die Betrachtung der Vertexnormalen-Hotspots für



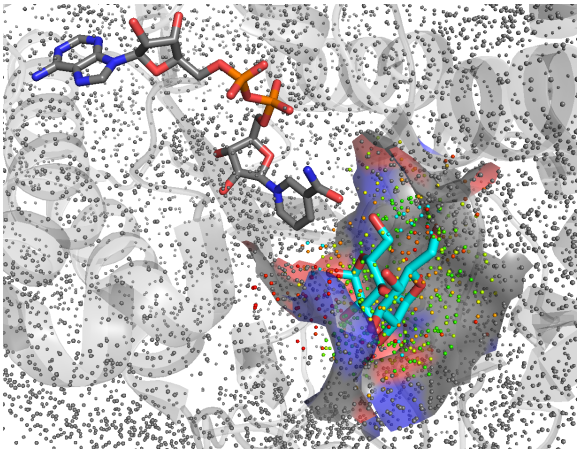
(a) 130 Hotspots des 1ANF.A-Epitops.



(b) Gematchten 2Z8D.A- und 1VL0.A-Hotspots.



(c) Gematchten 2Z8D.A-Hotspots.



(d) Gematchten 1VL0.A-Hotspots.

Abbildung 60: Gematchten Vertexnormalen-Hotspots.

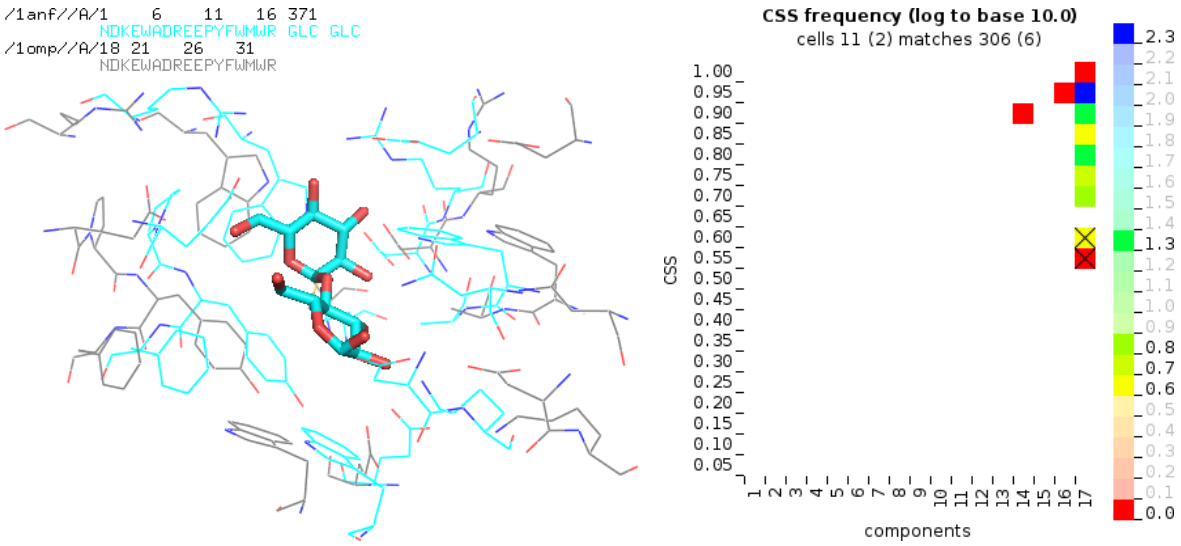
die Beschreibung der faltungsmusterunabhängigen Ähnlichkeit besser geeignet. Die Ähnlichkeit der beiden Epitope zu dem 1ANF.A-Epitop ist gegeben, sodass für die unidirektional, hypothetische Zuordnung der GO-Terme die folgende Aussage getroffen werden kann: Die Maltose aus 1ANF.A wird mit einer 70.3-prozentigen Wahrscheinlichkeit von 2Z8D.A und mit einer 64.4-prozentigen Wahrscheinlichkeit von 1VL0.A gebunden.

Die Rechenzeit für die Suche nach einer Vertexnormalen-Wolke von 130 Vertexnormalen in einer Wolke von 7904 Vertexnormalen im Fall von 2Z8D.A bzw. 6169 Vertexnormalen im Fall von 1VL0.A, beträgt derzeit 139s bzw. 153s auf einem Xeon X5650-Kern. Die geometrische Spezifität der Hotspots kann durch die höhere Dichte der Vertexnormalen gesteigert werden. Dies würde jedoch zu einer höheren Rechenintensität führen. Gleichzeitig könnten die Face-Typen CF, SF und TF berücksichtigt werden, sodass zusammen mit den 5 physiko-chemischen Eigenschaften $3 \cdot 5$ spezifische Hotspotklassen entstehen würden, die die erhöhte Rechenintensität aufgrund der höheren Vertexdichte wieder wettmachen würden. Die Koordinatenpaare der Vertexnormalen können zusätzlich mit den EP-Werten ausgestattet werden. Adaptive Poisson-Boltzmann Software (APBS) [19] ist ein Kontinuummodell für die Beschreibung des EP. Die mittels einer numerischen Lösung der Poisson-Boltzmann-Gleichung berechneten EPs der Oberflächen können an die Koordinaten der Vertexnormalen

als zusätzliche Eigenschaften gebunden werden, womit die Spezifität der Beschreibung des Epitops nochmals steigen würde. Die Implementierung der entsprechenden Schnittstelle zu APBS ist derzeit in Planung. Auch das Matchen der positiven und der negativen Isooberflächen ist möglich.

2.3.1.7 *Pseudoepitope*

Jede Struktur aus der PDB ist ein statischer Schnappschuss eines dynamischen Systems. Maltose-Epitope der Strukturen 1ANF.A und 1OMP.A sind Schnappschüsse der möglichen Holo- und Apo-Konformationen. Der einzige Unterschied zwischen den beiden Strukturen ist die Geometrie der Konformation. Abb. 61a zeigt die beiden Konformationsepitope nach



(a) Kombination des 1ANF.A- und des 1OMP.A-Epitops.

(b) Vgl. Abb. 54a.

Abbildung 61: Die Suche nach dem Konformationspseudoepitop mit den Alternativpositionen des Apo- und Holo-Epitops (a) führt zu einer deutlich besseren Erkennung der restlichen Epitope (vergleiche (b) mit Abb. 54a) und somit zu einer höheren Sensitivität.

ihrer Überlagerung. Jede Aminosäure des Holo-Epitops (cyan) korrespondiert mit einer identischen Aminosäure des Apo-Epitops (grau). Jede der 17 Aminosäuren des Konformationspseudoepitops wird somit durch zwei identische Aminosäuren mit den alternativen geometrischen Positionen repräsentiert. EpitopeMatch entscheidet im Matching-Prozess, welche der beiden Alternativen der jeweiligen Aminosäuren zum besseren CS auf der TS führt. Die Konformationspseudoepitope mit den Alternativpositionen sprechen unmittelbar das Induced-Fit-Problem an und erhöhen gleichzeitig die Sensitivität. Die Suche nach dem Konformationspseudoepitop auf 306 homologen Strukturen aus der NRPDB-Gruppe 211 resultiert in einem deutlich besseren Ergebnis (Abb. 61b im Vergleich zu Abb. 54a). Die Matches teilen sich nicht mehr in einen Holo- und einen Apo-Cluster (Abb. 54a), sondern befinden sich im Bereich $0.7 \leq CSS \leq 1.0$ (Abb. 61b), wobei die überwiegende Anzahl der Matches $CSS \geq 0.8$ ist und die FN-Epitope nach wie vor im Bereich $CSS < 0.6$ zu finden sind. Darüberhinaus ist die Anzahl der vollständig gematchten Epitope und TP-Epitope gestiegen. Von den 7 FN-Epitopen konnten 4 (4QRZ.A, 4QSD.A, 4QSE.A und 4QSE.B) als TP identifiziert werden, sodass die Sensitivität auf

$$sensitivity = \frac{TP}{TP + FN} = \frac{303}{303 + 3} \approx 0.99$$

gestiegen ist. Die konformationelle Spezifikation des auf 2 Repräsentativen (Apo- und Holo-Konformation) basierenden Pseudoepitops kann durch weitere Konformationsrepräsentativen aus der NRPDB-Gruppe 211 ergänzt werden um die Erkennung noch unempfindlicher gegenüber Induced-Fit zu gestalten. Darüber hinaus, existiert z.B. eine Trajektorie aus einer Molekulardynamiksimulation eines Epitops, so impliziert das Matchen der Trajektorien-Schnappschüsse die Messung der strukturellen und der dynamischen Ähnlichkeit.

Eine andere Variante der Pseudoepitope spricht das Problem der Multispezifität bzw. der Multimodalität der Bindung an. SiteEngines [160] untersucht das ATP-Epitop eines hypothetischen Proteins MJ0577 (1MJH). Die Autoren erstellen eine Binding Sites Data Base (BSDB) und repräsentieren ihren Inhalt und die PDB als Hotspots. Daraus ergeben sich drei Anwendungsarten: die Suche nach einem Epitop aus BSDB in der gesamten PDB; die Suche nach einem Epitop aus BSDB in der gesamten BSDB; und die Suche nach allen Epitopen aus der BSDB auf einer Struktur aus der PDB. Das Ziel ist die Erkennung der jeweiligen Funktion und somit die Unterstützung der Strukturgenomik- bzw. -proteomikprojekte. Jeder Anwendungsart liegen paarweisen Vergleiche der statischen Schnappschüsse zugrunde. So ergibt die Suche nach den ATP-Epitopen auf der vollständigen 1MJH eine Ähnlichkeit von $Matchscore = 44$ zu dem ATP-Epitop der 1ATP, und die Suche nach dem ATP-Epitop der 1MJH in der gesamten BSDB eine Ähnlichkeit von $Matchscore = 35$ zu dem ATP-Epitop der 1ATP. Demnach gibt es eine 44-prozentige Ähnlichkeit zwischen der ATP-Bindungsart der 1ATP und der 1MJH. Das bedeutet, dass das ATP an unterschiedliche Epitope und somit multispezifisch bindet.

Die PDB enthält gegenwärtig 1800 ATP-Protein-Komplexe. 1732 Ketten interagieren mit dem vollständigen (25 schwere Atome) ATP. 1616 Ketten sind einer bestimmten NRPDB-Gruppe zugeordnet und auf insgesamt 230 nicht redundante NRPDB-Gruppen verteilt. Die 6 homologen Ketten 1MJH.A-B und 3HGM.A-D gehören zu der NRPDB-Gruppe 2965, die mit $\approx 0.0035\%$ zu den am seltensten aufgeklärten ATP-Komplexen gehören. Pro NRPDB-Gruppe existiert genau eine repräsentative Struktur, die gleichzeitig den Rang 1 in der Gruppe besitzt. Die die repräsentative Struktur ist allerdings nicht unbedingt eine Holo-Struktur. Eine Holo-Repräsentative der jeweiligen Gruppe ist ein ATP-Komplex mit dem höchsten Rang innerhalb einer Gruppe. Es gibt also 230 nicht homologe ATP-Epitope. Das niedrigste Redundanzlevel lässt jedoch viele Mutationen und große Konformationsänderungen zu, sodass viele unterschiedliche Ligand-Epitop-Konformationen übersehen werden. Das höchste Redundanzlevel der NRPDB, in dem die Gruppen nur identische Ketten enthalten, erhöht die Anzahl der unterschiedlichen ATP-Komplexe auf 577. EpitopeMatch führt ein Kreuzvergleich der ATP-Konformationen durch. Abb. 62 zeigt die resultierende RMSD-Matrix. Die Matrix ist symmetrisch und paarweise-komplett nach der Pearson-Korrelation geclustert [46]. Die jeweiligen α-, β- und γ-Phosphatsauerstoffe der ATP-Moleküle sind vor der Überlagerung als geometrische Zentren zusammengefasst worden. Die RMSD-Werte verteilen sich zwischen $0.0\mathring{A} \leq RMSD < 3.75\mathring{A}$. Die maximale RMSD von $\approx 3.748\mathring{A}$ herrscht zwischen dem ATP.A.800 der 1I7L.A und dem ATP.A.1666 der 4A8F.A (Abb. 63a, sowohl das Adenosin, als auch das Triphosphat der ATP-Moleküle besitzen stark abweichende Konformationen). Das Konformationsspektrum des ATP lässt sich in ca. 5 große Cluster unterteilen, wobei die RMSD-Werte innerhalb eines Clusters mit $0.0\mathring{A} \leq RMSD \leq 1.0\mathring{A}$ verteilt sind. Eine ATP-Konformation mit der kleinsten RMSD zu dem Mittelwert aller ATP-Konformationen eines Clusters ist die ATP-Repräsentative, die ein Bindungsmodus des ATP-Moleküls repräsentiert, sodass das Vorhandensein mehrerer Cluster die Multimodalität der Bindung des ATP-Moleküls veranschaulicht. Pro Cluster existieren N nicht identische $5.0\mathring{A}$-Umgebungen, in die die Seitenketten der jeweiligen Proteinstruktur mit mindestens einem Atom hineinragen. Abb. 62 (schwarzer Pfeil links) zeigt die Lage des ATP.B.3001 der 1MJH.B im ATP-

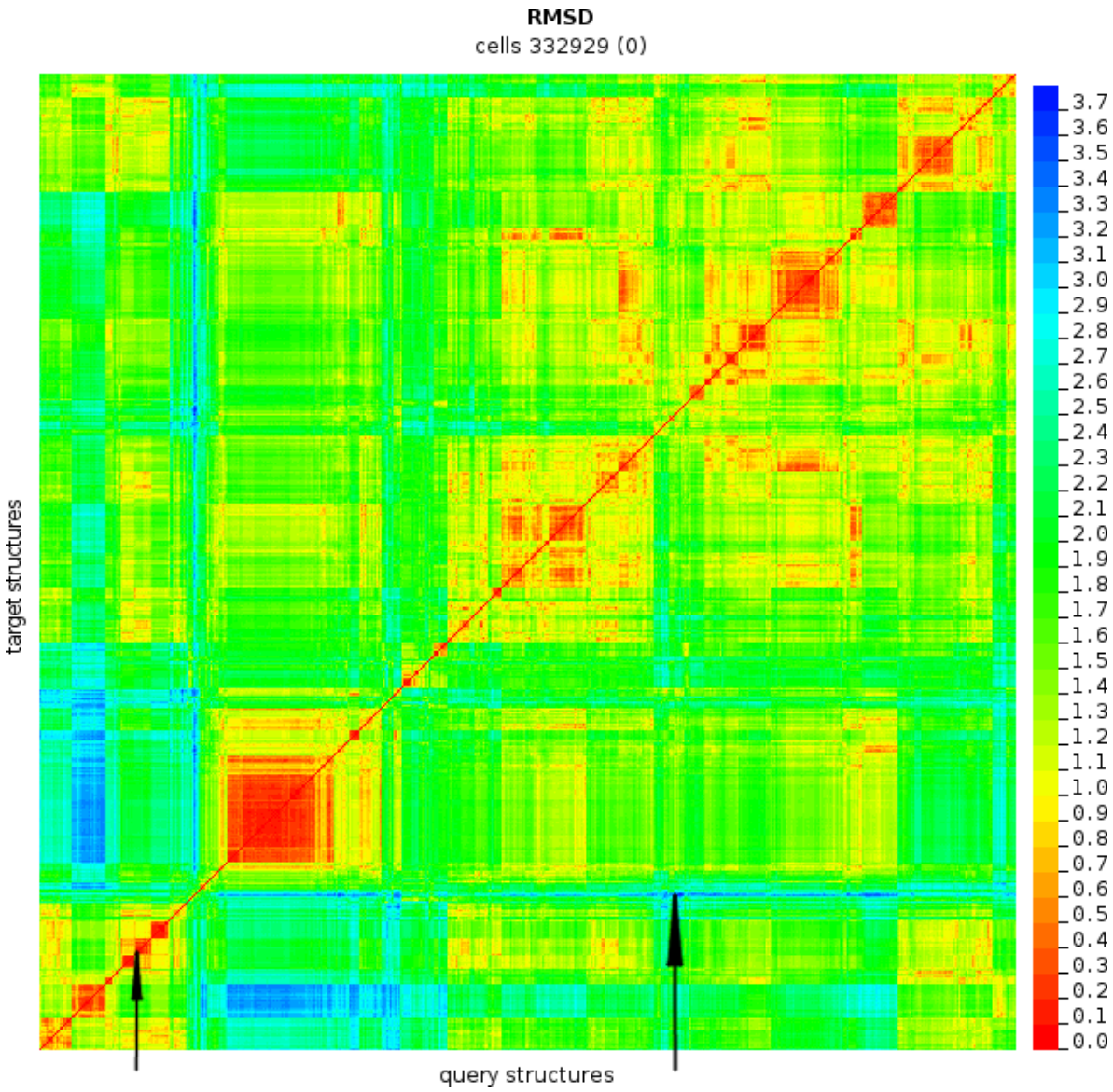
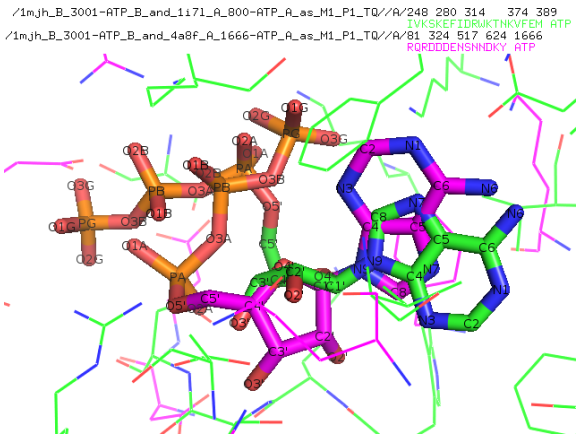
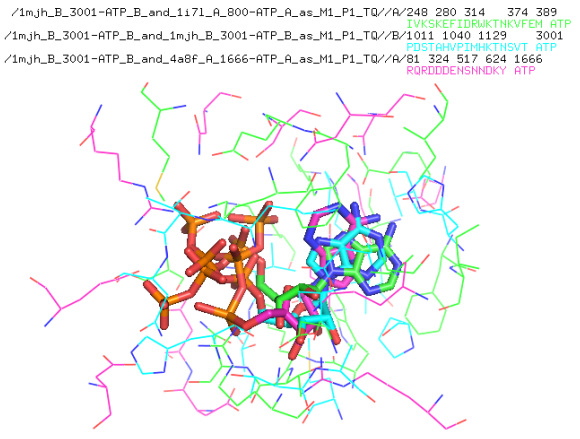


Abbildung 62: Vergleich der ATP-Konformationen aus 577 ATP-Komplexen. Alle Proteinstrukturen der Komplexe besitzen unterschiedliche Primärstrukturen. Das Konformationsspektrum des ATP lässt sich in ca. 5 große Cluster (verfolge die Diagonale) unterteilen, mit einer $0.0\mathring{A} \leq RMSD \leq 1.0\mathring{A}$ innerhalb des jeweiligen Clusters. Das Farbspektrum kodiert die RMSD. Der schwarze Pfeil links zeigt die Lage des ATP.B.3001 der 1MJH.B. Der schwarze Pfeil rechts zeigt die maximale $RMSD \approx 3.748\mathring{A}$ zwischen dem ATP.A.800 der 1I7L.A und dem ATP.A.1666 der 4A8F.A.

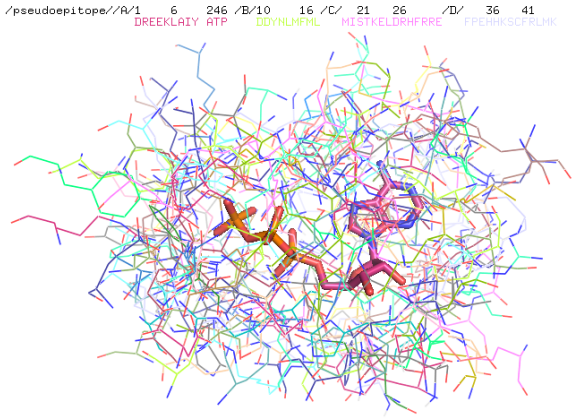
Konformationsraum. Dieses liegt mit $RMSD = 2.271\mathring{A}$ zum ATP.A.800 der 1I7L.A und mit $RMSD = 2.483\mathring{A}$ zum ATP.A.1666 der 4A8F.A relativ weit von den beiden Randkonformationen entfernt (Abb. 63b), die in den Zwischenräumen der Cluster liegen (Abb. 62 (schwarzer Pfeil rechts)). Angenommen, die Konformation des ATP.B.3001 der 1MJH.B ist die ATP-Repräsentative des Clusters. Dann bilden 47 ATP-Epitope mit mit $0.0\mathring{A} \leq RMSD \leq 1.0\mathring{A}$ zum ATP.B.3001 der 1MJH.B einen Cluster. Epitop des ATP.B.3001 der 1MJH.B selbst, und des ATP.A.148 der homologen Struktur 3HGM.A (NRPDB-Gruppe 2965 des niedrigsten Redundanzlevels) werden ausgeschlossen. Die verbleibenden 45 Epitope werden anhand der ATP-Moleküle mit dem Epitop des ATP.B.3001 der 1MJH.B überlagert. Redundante Aminosäurenpositionen werden ausgeschlossen, indem alle Aminosäuren paarweise miteinan-



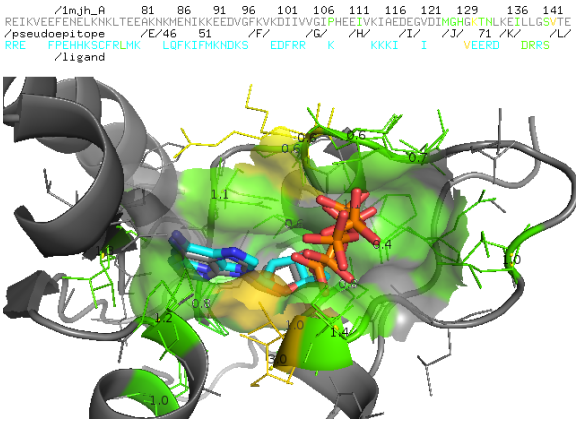
(a) Maximaler Konformationsunterschied.



(b) Konformationslage des 1MJH.B-ATP.



(c) ATP-Pseudoepitop mit 245 Aminosäuren.



(d) CS aus 17 Aminosäuren der 1MJH.A.

Abbildung 63: ATP-Pseudoepitop. Im Gegensatz zu einem Konformationspseudoepitop (vgl. Abb. 61a), das pro Residuum mehrere Alternativpositionen enthält und sich aus dem gleichen Epitop mit den unterschiedlichen Konformationen zusammensetzt, setzt sich ein Pseudoepitop aus den Residuen der unterschiedlichen Epitope zusammen, die nach der Konformation des Liganden überlagert sind.

der verglichen werden. Wenn ein Aminosäurenpaar eine $NWRMSD > 0.5$ besitzt, dann wird eine Aminosäure mit der niedrigeren physiko-chemischen Priorität ausgeschlossen. Die Prioritäten sind nach den 10 physiko-chemischen Eigenschaften [106] festgelegt, sodass die geladenen Aminosäuren die höchste Priorität besitzen, gefolgt von den aromatischen, polaren, aliphatischen, etc.. Auf diese Weise entsteht ein Pseudoepitop für eine bestimmte ATP-Konformation. Abb. 63c zeigt das resultierende Pseudoepitop mit insgesamt 245 Aminosäuren. Diese verteilen sich auf insgesamt 39 Pseudochains, von denen jede einem der 45 Epitope entspricht. 6 Epitope sind aufgrund der redundanten Positionen ihrer Aminosäuren vollständig ausgeschlossen. Das ATP-Epitop der 1MJH.B besteht aus 17 Aminosäuren (Abb. 63b, cyan, 5.0$\AA$-Umgebung). Ergebnis der Suche nach dem Pseudoepitop (Abb. 63c) auf 1MJH.A ist in der Abb. 63d dargestellt. Es handelt sich um eine 17 Aminosäuren große CS mit $IDENT = 0.412$, $SSIM = 0.92$, $RMSD = 1.474\AA$ und $CSS = 0.906$, die sich über 8 Pseudochains mit 1, 1, 3, 5, 1, 3, 2 und 1 Aminosäuren des Pseudoepitops verteilt. Die ATP-Repräsentative des Pseudoepitop-Clusters und das ATP der 1MJH.A sind nahezu deckungsgleich Überlagert. Die Anzahl der Kollisionen der mittransformierten ATP-Repräsentative mit den Aminosäuren der 1MJH.A ist gleich 0. Während SiteEngines eine maximal 44-

prozentige Wahrscheinlichkeit für eine ATP-Bindungsfunktion auf 1MJH.A erkennt, sagt EpitopeMatch mittels der Pseudoepitop-Methode eine 90.6-prozentige Wahrscheinlichkeit voraus.

2.3.2 *Biopolymere*

Das bisher beschriebene Konzept ist anwendbar auf beliebige Biopolymere aus der PDB und ihre Komponenten aus der CCD.

2.3.2.1 *DNA*

Click [123] demonstriert auf dem Webserver den Vergleich zweier DNA-Doppelhelices 1YSA.A-B und 2AYG.C-D. Dabei deklariert es die C3'-Atome der Nukleoside als Repräsentativen. EpitopeMatch wiederholt den Vergleich der Doppelhelices anhand der C3'-Atome. Abb. 65a zeigt eine CS mit 27 korrespondierenden C3'-Atomen, $IDENT = 0.185$ (entspricht $27 \cdot 0.185 = 5$ identischen Nukleotiden), $RMSD = 1.728Å$ und $MCSS = 0.665$. Abgesehen von einem korrespondierenden Paar (1YSA.A.DT.19; 2AYG.C.DG.14), dass zu einer besseren RMSD führt (abweichendes Paar von Click ist (1YSA.A.DT.19; 2AYG.C.DG.13), resultiert in $RMSD = 1.8Å$), ist das Alignment identisch. Das Phosphatdesoxyribose-Rückgrat der DNA

4	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
DA 1	TYPE	OP3	P	OP1	OP2	O5'	C5'	C4'	O4'	C3'	O3'	C2'	C1'	N9	C8	N7	C5	C6	N6	N1	C2	N3	C4	H0P3	H0
DA 1	1	1	2	3	4	5	6	7	8	9	10	11	12												
DC 1	TYPE	OP3	P	OP1	OP2	O5'	C5'	C4'	O4'	C3'	O3'	C2'	C1'	N1	C2	O2	N3	C4	N4	C5	C6	H0P3	H0P2	H5'	H5
DC 1	1	1	2	3	4	5	6	7	8	9	10	11	12												
DG 1	TYPE	OP3	P	OP1	OP2	O5'	C5'	C4'	O4'	C3'	O3'	C2'	C1'	N9	C8	N7	C5	C6	O6	N1	C2	N2	N3	C4	H0
DG 1	1	1	2	3	4	5	6	7	8	9	10	11	12												
DT 1	TYPE	OP3	P	OP1	OP2	O5'	C5'	C4'	O4'	C3'	O3'	C2'	C1'	N1	C2	O2	N3	C4	O4	C5	C7	C6	H0P3	H0P2	H5
DT 1	1	1	2	3	4	5	6	7	8	9	10	11	12												

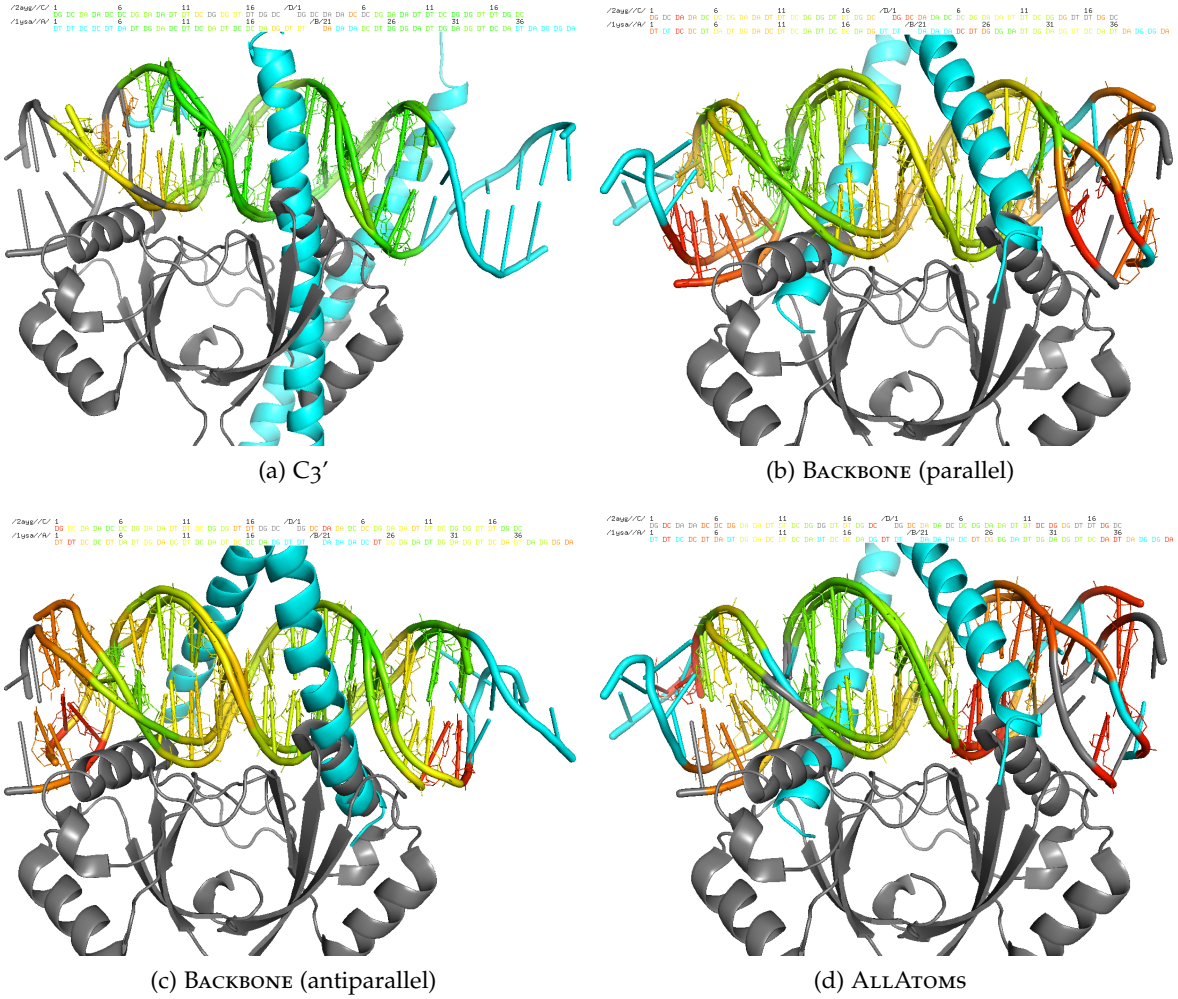
(a) Backbone. Sowohl Purine als auch Pyrimidine sind vom gleichen Typ (unspezifisch).

4	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
DA 1	TYPE	OP3	P	OP1	OP2	O5'	C5'	C4'	O4'	C3'	O3'	C2'	C1'	N9	C8	N7	C5	C6	N6	N1	C2	N3	C4	H0P3	H0
DA 1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22		
DC 1	TYPE	OP3	P	OP1	OP2	O5'	C5'	C4'	O4'	C3'	O3'	C2'	C1'	N1	C2	O2	N3	C4	N4	C5	C6	H0P3	H0P2	H5'	H5
DC 1	2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20				
DG 1	TYPE	OP3	P	OP1	OP2	O5'	C5'	C4'	O4'	C3'	O3'	C2'	C1'	N9	C8	N7	C5	C6	O6	N1	C2	N2	N3	C4	H0
DG 1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		21	22	
DT 1	TYPE	OP3	P	OP1	OP2	O5'	C5'	C4'	O4'	C3'	O3'	C2'	C1'	N1	C2	O2	N3	C4	O4	C5	C7	C6	H0P3	H0P2	H5
DT 1	2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19		20			

(b) AllAtoms. Purine sind vom Typ 1 und Pyrimidine sind vom Typ 2 (spezifisch).

Abbildung 64: DNA-Templates. Neben den Angaben der Distanzmatrixnummern pro Atom wird pro Residuum der Typ (Spalte 1) festgelegt. Anhand der Typ-Angaben wird im Fall (a) eine unspezifische und im Fall (b) eine spezifische Substitutionsmatrix erzeugt.

besteht aus 12 schweren Atomen. Abb. 64a zeigt ihre Verteilung auf 12 Korrespondenzlevels (Distanzmatrizen), sodass die gesamte geometrische Information des DNA-Rückgrats berücksichtigt wird. Die Korrespondenzen der C3'-Atome werden in diesem Fall von der neunten Distanzmatrix erfasst. Da die Geometrie der Basen außer Acht gelassen ist, sind die Nukleotidtypen vorerst unwichtig und mit $TYPE = 1$ als unspezifisch definiert. Die Suche nach der vollständigen Rückgratgeometrie liefert eine CS aus 32 Phosphatdesoxyribosen mit $IDENT = 0.188$ (entspricht $32 \cdot 0.188 = 6$ identischen Nukleotiden), $RMSD = 3.036Å$ und $MCSS = 0.62$ (Abb. 65b), und eine CS aus 33 Phosphatdesoxyribosen mit $IDENT = 0.121$ (entspricht $33 \cdot 0.121 = 4$ identischen Nukleotiden), $RMSD = 2.763Å$ und $MCSS = 0.677$ (Abb. 65c). Der wesentliche Unterschied zwischen den beiden Matches ist die Alignmentausrichtung der Doppelhelices, die im ersten Fall beide von 5' nach 3' verlaufen und im zweiten Fall antiparallel, eine Doppelhelix von 5' nach 3' überlagert mit der zweiten von 3' nach 5'. Abb. 64b zeigt ein Template, in dem neben den Rückgratatomen auch die Ba-



(a) C3′ (b) BACKBONE (parallel)

(c) BACKBONE (antiparallel) (d) ALLATOMS

Abbildung 65: Auch die DNA-Matching-Ergebnisse sind abhängig von der Wahl der Matching-Templates und somit von der Wahl der Auflösung des Informationsgehalts.

senatome berücksichtigt werden. Dabei verteilen sich die Atome der Purine (DA oder DG, $TYPE = 1$) über 22, und die Atome der Pyrimidine (DC oder DT, $TYPE = 2$) über 20 Korrespondenzlevels. Die unterschiedlichen Typen definieren eine Spezifität im Bezug auf die Unterscheidung zwischen den Purinen und Pyrimidinen, wobei die beiden Purine und die beiden Pyrimidine im einzelnen einander entsprechen dürfen. Anhand dieses Templates gefundene CS besteht aus 28 korrespondierenden Nukleotiden mit $IDENT = 0.286$ (entspricht $28 \cdot 0.286 = 8$ identischen Nukleotiden), $RMSD = 3.6\text{Å}$ und $MCSS = 0.515$ (Abb. 65d). Die CSs aus Abb. 65b und Abb. 65d haben einen gemeinsamen Kern aus 16 Nukleotiden. Das führt zu einer sehr ähnlichen Überlagerung der 1YSA und 2AYG. Die CS aus Abb. 65d ist kleiner, da nur die Kombinationen AA, AG, GG und CC, CT, TT zulässig sind. Sowohl das BACKBONE-Template mit der vollständigen Rückgratgeometrie als auch das ALLATOMS-Template mit der vollständigen Nukleotidgeometrie mit der Berücksichtigung der Purin-Pyrimidin-Verteilung liefern ein deutlich aussagekräftigeres Ergebnis als das nackte C3′-Template.

Die Kernaussage dieser Demonstration ist, dass weder die alleinige Betrachtung der Cα-Atome der Proteine, noch die alleinige Betrachtung der C3′-Atome der DNA für die vollständigen Strukturen repräsentativ ist. Die Wahl der repräsentativen Atomgruppen muss immer im Sinne der Auflösung des erforderlichen Informationsgehalts erfolgen.

2.4 BENCHMARK

Es existiert eine ganze Reihe von Benchmarks für paarweises Alignment für die SCA-Algorithmen. Der Repetitions, Indels, Permutation and Conformational variability (RIPC)-Datensatz gilt als das anspruchsvollste Benchmark in diesem Feld [113]. Die Autoren stellen im Einzelnen fest, dass die unterschiedlichen Algorithmen beim Vergleich von entfernt homologen Strukturen unterschiedliche Alignments produzieren und nicht immer den Referenzalignments entsprechen. Sie rufen die Entwickler dazu auf, ihre Methoden zu verbessern.

2.4.1 *Datensatz & Ergebnisse*

Die gestellte Herausforderung des RIPC-Benchmarks an die SCA-Algorithmen besteht in den unterschiedlichen Typen der Strukturveränderungseigenschaften. Die Repetitionen und die IDs nehmen Einfluss auf die Länge der Sequenzen bzw. die Größe der Strukturen. Die CPs sorgen für eine Veränderung der Topologie bzw. der SSE-Konnektivität bezüglich der Syntheserichtung und zerstören somit die Kontinuität der Sequenzordnung. Alle drei können neben Induced-Fit zu den beträchtlichen Konformationsänderungen führen, die in vielen Fällen nur flexibel gematcht werden können. Ein SCA-Algorithmus muss also bestenfalls mit allen Schwierigkeiten umgehen können, um deren Bewältigung nicht dem Benutzer durch das Verteilen der Aufgabe auf unterschiedliche Methoden zu überlassen, sondern um deren Aufschlüsselung so plausibel wie möglich an den Benutzer heranzutragen.

RIPC-Datensatz besteht aus 40 Domänenpaaren. Für 23 Domänenpaare sind Referenzalignments angegeben, die auf den Alignments zu den homologen Strukturen, auf der Residuennummerierung der PDB oder auf der Suche nach den Residuen mit der äquivalenten Funktion basieren. Tab. 15 enthält 23 Domänenpaare, die entweder eine oder zwei der oben beschriebenen Strukturveränderungseigenschaften aufweisen. Neben den Domänengrößen ist pro Domänenpaar die Seuquenzidentität angegeben. Für jedes Domänenpaar existiert ein Referenzalignment, an dem die Alignmentkonsistenz der jeweiligen SCA-Methode gemessen wird. Ergebnisse von insgesamt 12 als "state of the art" geltenden Algorithmen sind aus [113, 145, 41] zusammengetragen. Ergebnisse von MULTIPROT sind mittels Webserver ergänzt. Ergebnisse von CLICK sind mittels Webserver ermittelt. EPITOPEMATCH stellt 6

geom. / phys.-chem.	Cα (CA), niedrig	BACKBONE (BB), mittel	ALLATOMS (AA), hoch
UNSPECIFIC (U), keine	am niedrigsten	mittel	hoch
WEIGHTED (W), hoch	niedrig +	mittel +	am höchsten

Tabelle 14: Informationsgehalt der Template-Kombinationen.

Matching-Varianten zur Verfügung, die aus der Kombination der drei geometrischen Templates Cα (CA), BACKBONE (BB), ALLATOMS (AA) mit den 2 physiko-chemischen Templates UNSPECIFIC (U) und WEIGHTED (W) resultieren (Tab. 14). Somit reicht das Matching-Spektrum von der puren Geometrie der Cα-Atome im Fall von CA/U bis zur vollständigen Geometrie, gewichtet mit AVEHYDROP , MOLWEIGHT und NETCHARGE der Aminosäuren im Fall von AA/W. 13 Domänenpaare enthalten keine CPs (Tab. 15, P) und können in der Regel mit den kontinuierlichen SCA-Algorithmen gematcht werden. Die restlichen 10 Domänenpaare im unteren Teil der Tabelle enthalten CPs und können in der Regel nur mit den diskontinuierlichen SCA-Algorithmen gematcht werden. Für jede SCA-Methode und Domänenpaar ist der Prozentsatz (gerundet) des erkannten Referenzalignments angegeben. Diese Werte sind für jede SCA-Methode für den kontinuierlichen Block, diskontinuierlichen Block und

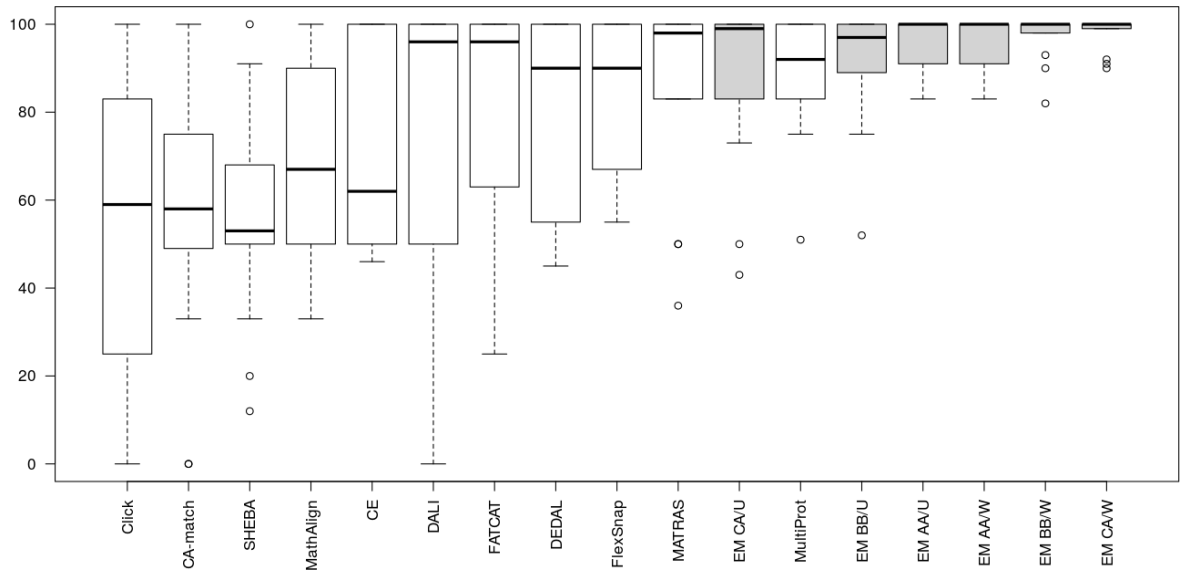
Domain 1	Domain 2	Type	Size 1	Size 2	Seq. Id.	Ref. Size	MatAlign	SHEBA	Click	CE	MATRAS	FATCAT	Cα-match
d1dlia1	d1mv8a1	C	98	98	26.3	4	50	100	100	50	50	100	50
d1ggga_	d1wdna_	C	220	223	100	220	99	68	59	53	98	100	58
d1l5ba_	d1l5ea_	C	101	101	100	101	50	53	93	60	100	100	49
d2bbma_	d4clna_	C	148	148	100	148	47	91	99	50	95	96	0
d1d5fa_	d1nd7a_	C+I	350	374	34.6	6	67	50	33	66	83	100	50
d1hava_	d1kxfa_	C+I	216	159	20.1	4	100	50	0	100	100	25	100
d1jj7a_	d1lvga_	C+I	251	190	21.1	8	50	12	0	62	100	100	0
d1an9a1	d1npxa1	I	247	198	19.3	11	73	90	46	46	36	63	90
d1ay9b_	d1b12a_	I	108	239	20	10	90	20	80	100	90	90	90
d1crla_	d1edea_	I	534	310	22	3	67	67	67	100	100	67	33
d1gbga_	d1ovwa_	I	214	398	19.5	3	33	33	0	67	100	33	67
d1hcya2	d1lnlb1	I	263	307	13.9	4	75	50	25	50	50	50	75
d2adma_	d2hmyb_	I	386	327	15.3	12	100	67	83	100	100	100	58
							69.3	57.8	52.7	69.5	84.8	78.8	55.4
d1nkla_	d1qdma1	P	78	77	24.3	72	0	0	0	0	0	0	41
d1nlsa_	d2bqpa_	P	237	228	43.9	6	0	100	83	83	0	83	83
d1qasa2	d1rsya_	P	126	135	26.4	75	84	80	99	84	84	84	94
d1gsaa1	d2hgsa1	P+C	122	102	15.6	5	20	40	40	40	40	40	40
d1jwyb_	d1puja_	P+I	281	261	20.2	12	0	0	67	33	0	0	91
d1jwyb_	d1uola2	P+I	281	212	19.5	11	0	0	82	36	0	0	90
d1kiaa_	d1nw5a_	P+I	275	270	19.9	12	0	0	83	8	16	25	67
d1nw5a_	d2adma_	P+I	270	386	16.4	13	0	46	39	15	46	46	76
d1qq5a_	d3chya_	P+I	245	128	20.2	3	0	0	0	0	0	0	0
d1b5ta_	d1k87a2	P+R	275	351	17.5	8	0	0	0	63	0	0	0
							10.4	26.6	39.4	36.2	18.6	27.8	58.2
							43.7	44.2	46.9	55	56	56.6	56.6

(a) Teil 1.

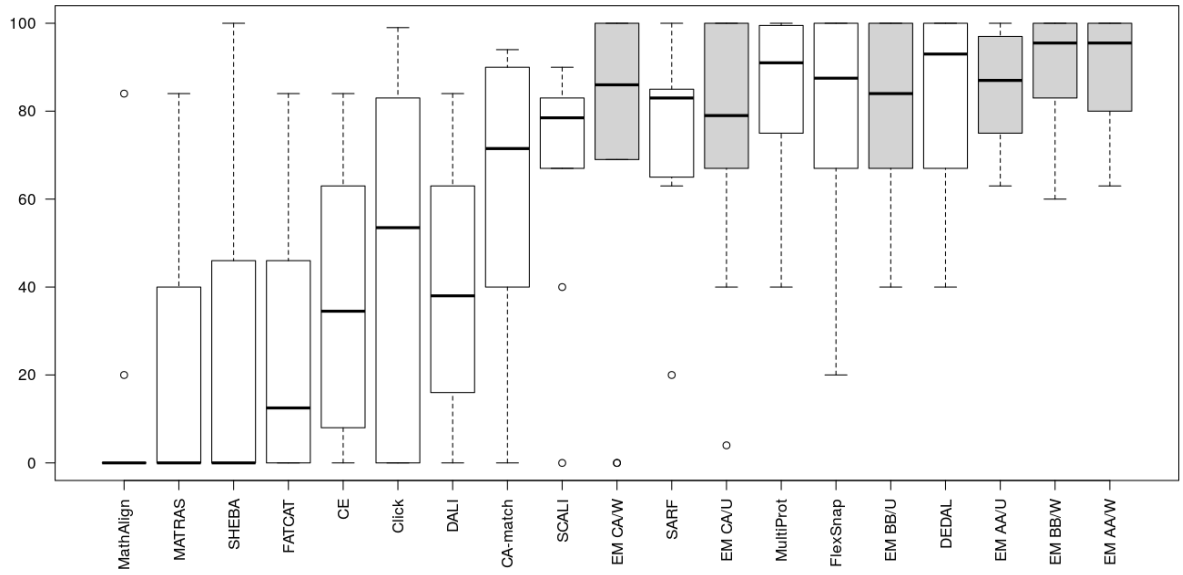
Domain 1	Domain 2	SCALI	SARF	DALI	DEDAL	EM CA/U	FlexSnap	MultiProt	EM BB/U	EM CA/W	EM AA/U	EM BB/W	EM AA/W
d1dlia1	d1mv8a1			50	50	50	100	100	100	100	100	100	100
d1ggga_	d1wdna_			96	96	96	100	51	97	99	98	98	98
d1l5ba_	d1l5ea_			50	99	99	100	99	52	99	100	100	100
d2bbma_	d4clna_			0	45	43	93	78	89	92	88	93	89
d1d5fa_	d1nd7a_			50	83	83	67	83	83	100	83	100	83
d1hava_	d1kxfa_			100	100	100	100	100	100	100	100	100	100
d1jj7a_	d1lvga_			100	100	100	100	100	100	100	100	100	100
d1an9a1	d1npxa1			100	55	73	55	91	91	91	91	82	91
d1ay9b_	d1b12a_			90	90	90	90	90	90	90	90	90	90
d1crla_	d1edea_			100	67	100	67	100	100	100	100	100	100
d1gbga_	d1ovwa_			100	100	100	67	100	100	100	100	100	100
d1hcya2	d1lnlb1			50	50	100	75	75	75	100	100	100	100
d2adma_	d2hmyb_			100	100	100	83	92	100	100	100	100	100
				75.9	79.6	87.2	84.4	89.2	90.5	97.8	96.2	97.2	96.2
d1nkla_	d1qdma1	69	92	0	94	4	100	68	85	69	94	90	92
d1nlsa_	d2bqpa_	83	83	83	100	100	83	100	100	100	100	100	100
d1qasa2	d1rsya_	82	65	84	97	99	100	93	97	99	97	99	99
d1gsaa1	d2hgsa1	40	20	40	40	40	20	40	40	80	80	60	80
d1jwyb_	d1puja_	83	83	33	92	83	92	92	67	92	92	92	92
d1jwyb_	d1uola2	90	100	36	100	100	100	91	100	100	82	100	100
d1kiaa_	d1nw5a_	75	83	16	83	75	75	75	83	75	75	83	75
d1nw5a_	d2adma_	84	85	46	100	100	100	92	100	100	100	100	100
d1qq5a_	d3chya_	67	67	0	67	67	67	67	67	0	67	100	100
d1b5ta_	d1k87a2	0	63	63	50	75	50	50	63	0	63	63	63
		67.3	74.1	40.1	82.3	74.3	78.7	76.8	80.2	71.5	85	88.7	90.1
				60.3	80.8	81.6	81.9	83.8	86	86.4	91.3	93.5	93.6

(b) Teil 2.

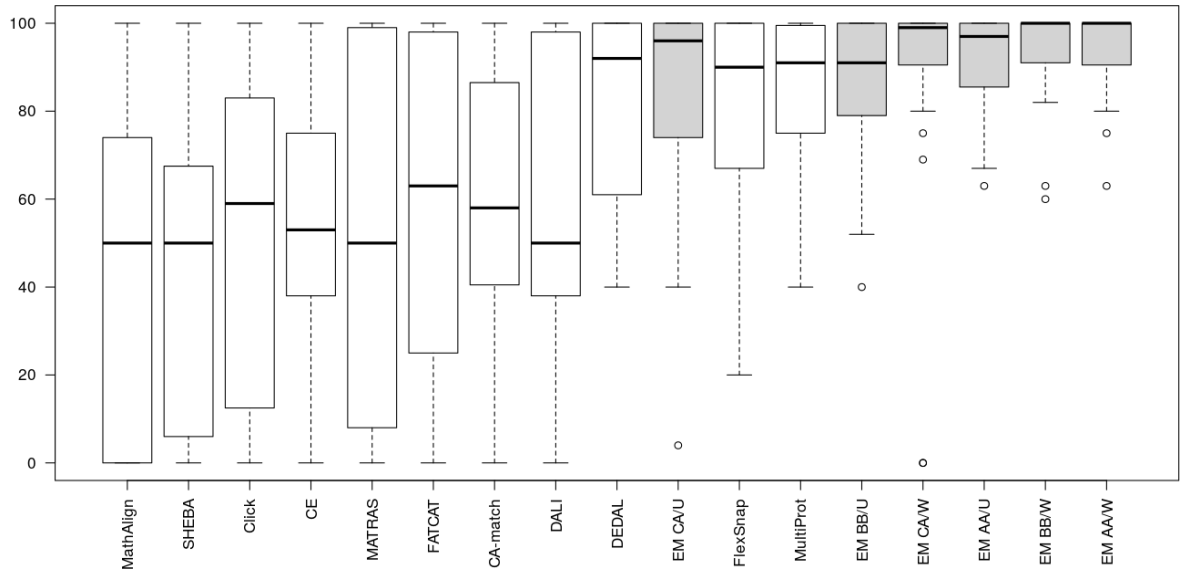
Tabelle 15: RIPC-Benchmark. Die Domänenpaare sind in zwei Blöcke geteilt: kontinuierlich (Typ C, I, C+I) und diskontinuierlich (Typ P, P+C, P+I, P+R). Pro Domänenpaar & Algorithmus ist der Prozentsatz der erkannten Referenzalignments angegeben. Die Prozentzahlen sind pro Block & Algorithmus, und insgesamt pro Algorithmus gemittelt (graue Zellen). Die Algorithmen sind nach dem Prozentsatz-Mittelwert der erkannten Referenzalignments von links nach rechts sortiert.



(a) Prozentsätze der erkannten Referenzalignments im kontinuierlichen Block (Typ C, I, C+I).



(b) Prozentsätze der erkannten Referenzalignments im diskontinuierlichen Block (Typ P, P+C, P+I, P+R).



(c) Prozentsätze der erkannten Referenzalignments insgesamt.

Abbildung 66: RIPC-Benchmark. Boxplot-Darstellung der Daten aus Tab. 15.

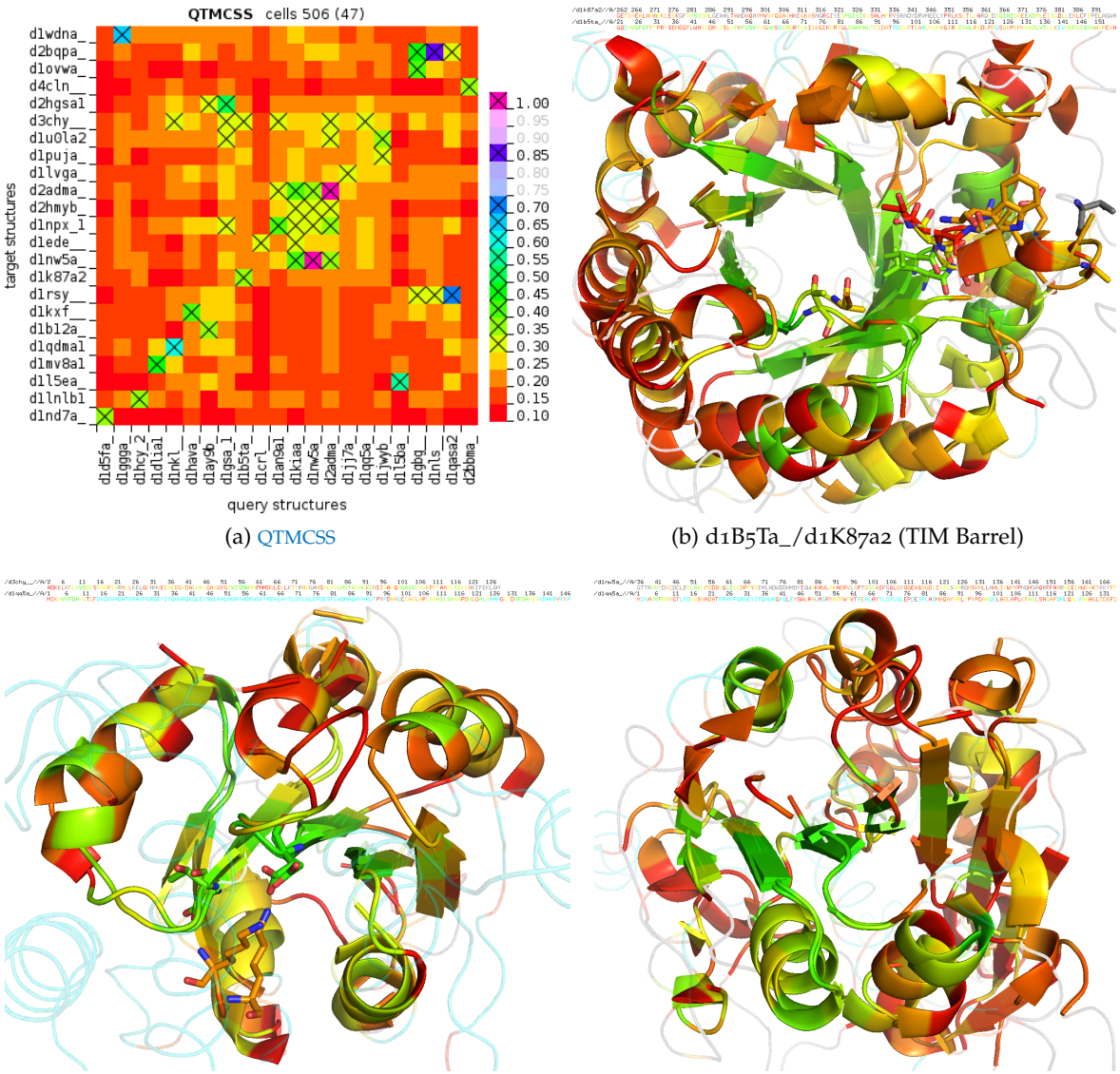
insgesamt gemittelt (grauer Hintergrund). Die Reihenfolge der SCA-Methoden entspricht ihrer aufsteigenden Sortierung nach den gemittelten Prozentsätzen aller Referenzalignments. SCALI und SARF fallen aus der Reihe, da keine Werte für den kontinuierlichen Block publiziert sind und mangels Verfügbarkeit der Webserver bzw. zum Download bereitgestellten Software nicht nachgerechnet werden konnten. Abb. 66 veranschaulicht die Werte aus Tab. 15 blockweise mit den Boxplots. Die Boxen der Matching-Varianten von EpitopeMatch sind grau. Die Parameter für alle 6 Matching-Varianten von EpitopeMatch sind abgesehen von unterschiedlichen Kombinationen der Templates gleich. Alle Domänenpaare sind im Modus *invariant* (bezüglich der Syntheserichtung) und *diskontinuierlich* gematcht.

Domänenpaare aus dem kontinuierlichen Block (Abb. 66a) enthalten IDs (I), Konformationsänderungen (C) und Kombination aus beiden (C+I). Mit dem geringsten Satz an Information (CA/U) liegt EpitopeMatch mit 87.2% erkannter Referenzalignments auf Augenhöhe mit MATRAS 84.8% und MultiProt 89.2%. Mit dem steigenden Informationsgehalt (Tab. 14) erkennt EpitopeMatch $>$ 90% der Referenzalignments. Der höchste Prozentsatz von 97.8% wird allerdings mit der CA/W-Template-Kombination (Informationsgehalt "niedrig +") erreicht. Eine mögliche Erklärung dafür ist die Vorgehensweise der Autoren [113] bei der Erstellung der Referenzalignments, indem sie sich nach der Cα-Superpositionierung gerichtet haben.

Der diskontinuierliche Block (Abb. 66b) ist von den CPs (P) geprägt. Die drei ersten Template-Kombinationen CA/U, CA/W und BB/U mit dem niedrigen Informationsgehalt liegen mit 74.3%, 71.5% und 80.2% im Feld von SARF 74.1%, MultiProt 76.8% und FlexSnap 78.7%. DEDAL erreicht mit 82.3% als einziger die Marke von $>$ 80%. Die Template-Kombinationen BB/W, AA/U und AA/W mit dem hohen Informationsgehalt liegen mit 88.7%, 85% und 90.1% deutlich vorne.

In der Summe (Abb. 66c) liegt EpitopeMatch mit CA/U 81.6% knapp hinter FlexSnap 81.9% und MultiProt 83.8%. Ergänzt man jedoch die nackte Cα-Geometrie um die weiteren Atome und um die physiko-chemischen Eigenschaften, so liegt EpitopeMatch angefangen mit BB/U 86% deutlich vorne. Die informationsreichste Matching-Variante AA/W erkennt mit 93.6% 9.8% mehr Referenzalignments als das MultiProt mit 83.8%. Während MATRAS, SARF, DEDAL, FlexSnap und MultiProt sich abhängig von den Strukturveränderungseigenschaften in der Konsistenz der Erkennung der Referenzalignments abwechseln, bleiben die Template-Kombinationen BB/W, AA/U und AA/W mit dem hohen Informationsgehalt konsistent, und mit einem Abstand an der Spitze. Abgesehen von Click erkennt keine der aufgeführten Methoden diskontinuierliche Epitope. Das Problem von Click ist jedoch eine hohe Empfindlichkeit gegenüber von Konformationsänderungen.

Die zum Vergleich gewählten Algorithmen arbeiten mit dem aus methodischen Gründen reduzierten Informationsgehalt. Der bewusste Ausschluss der vorhandenen Information im Interesse der Performance führt zu einem inakzeptablen Qualitätsverlust. Die richtige Interpretation der Ähnlichkeit ist jedoch ein Fundament für die Vergleichbarkeit, auf der die Musterklassifizierung basiert. Die Angabe der Anzahl, der Identität und der RMSD der korrespondierenden Residuen ist für die Beschreibung der Ähnlichkeit nicht ausreichend. Abb. 67a demonstriert den Kreuzvergleich der 23 Domänenpaare im Matching-Modus AA/W. Die Matrix ist unsymmetrisch und besteht aus 22 (QSs, Tab. 15 Spalte "Domain 1") mal 23 (TSs, Tab. 15 Spalte "Domain 2") Matches. Der gewählte Deskriptor QTMCSS bewertet die größte gemeinsame Substruktur im Bezug auf die Größe der QSs und der TSs und ist als ein Maß für die Gesamtähnlichkeit zweier Strukturen zu betrachten. Die TP-Matches befinden sich im Bereich $0.3 \leq QTMCSS \leq 0.9$. EpitopeMatch findet als einzige Methode das vollständige, aus drei korrespondierenden Aminosäuren (ASP.8.A/ASP.57.A, SER.114.A/THR.87.A und LYS.147.A/LYS.109.A) bestehende Referenz-



(a) QTMCSS

(b) d1B5Ta_/d1K87a2 (TIM Barrel)

(c) d1QQ5a_/d3CHYa_ (Rossmann fold)

(d) d1QQ5a_/d1NW5a_ (Rossmann fold)

Abbildung 67: RIPC-Kreuzvergleich. Kreuzvergleich ist das paarweise multiple Strukturalignment, das die Erkennung von gemeinsamen Substrukturen ermöglicht, nach denen die Strukturen klassifiziert werden können.

alignment des Domänenpaares d1QQ5a_/d3CHYa_. Abb. 67c zeigt die MCS der beiden Domänen. Die drei korrespondierenden Aminosäuren sind als *sticks* hervorgehoben. Deutlich zu sehen ist die ähnliche Ausrichtung der identischen Referenzresiduen, deren Erkennung durch die vollständige Geometrie und physiko-chemische Gewichtung ermöglicht wird. Die Matrix ist paarweise-einfach nach der Pearson-Korrelation geclustert [46]. Das Domänenpaar d1QQ5a_/d3CHYa_ markiert die rechte obere Ecke eines zentral gelegenen quadratischen (9x9) Clusters. Auch wenn die Gesamtähnlichkeit der meisten Paare innerhalb dieses Clusters $QTMCSS < 0.3$ ist, verfügen diese über ein gemeinsames Faltungsmuster, das nach CATH die Topologie "Rossmann fold" und die Architektur "3-Layer($\alpha\beta\alpha$) Sandwich" besitzt, und der Klasse "Alpha Beta" zugeordnet ist. Allerdings unterscheiden sich die entsprechenden homologen Superfamilien. Abb. 67d demonstriert die MCS des Domänenpaares d1QQ5a_/d1NW5a_ mit einer sehr geringen, aber dennoch erkennbaren Ähnlichkeit der besagten Architektur. Eine weitere Schwierigkeit beim Matchen entsteht, wenn neben einer

beträchtlichen Konformationsänderung eine gewisse Konformationssymmetrie vorhanden ist. Abb. 67b zeigt die MCS der "TIM Barrel" des Domänenpaares d1B5Ta_/d1K87a2. Die ringförmig angeordneten β-Faltblätter sind von den ebenfalls ringförmig angeordneten α-Helices umgeben, die sich jedoch stark in ihrer Konformation unterscheiden. In solchen Fällen werden mehrere Alignments generiert, die sich in ihrer Bewertung kaum voneinander unterscheiden. EpitopeMatch erkennt als einzige Methode 6 von 8 korrespondierenden Referenzresiduen, allerdings in diesem Fall anhand der CA/U-Template-Kombination mit dem niedrigsten Informationsgehalt. Die Tatsache, dass alle Methoden mit der Erkennung des vollständigen Referenzalignments dieses Domänenpaares Schwierigkeiten haben liegt womöglich in der falschen Definition der Referenzresiduen.

2.4.2 *Signifikanz*

Die Kreuzvergleichmatrix (Abb. 67a) enthält pro Domänenpaar die am höchsten bewertete MCS. 5 Domänenpaare d1DLIa1/d1MV8a1, d1GGGa_/d1WDNa_, d1L5Ba_/d1L5Ea_, d2BBMa_/d4CLN__ und d1D5Fa_/d1ND7a_ aus dem kontinuierlichen Block (Tab. 15) weisen jedoch so große Konformationsänderungen (C) auf, dass sie nur flexibel gematcht werden können. EpitopeMatch liefert in allen 5 Fällen jeweils 2 MCSs, die insgesamt möglichst vollständige gemeinsame Substruktur abdecken. Abb. 68 zeigt 22 Boxplots, ein Boxplot

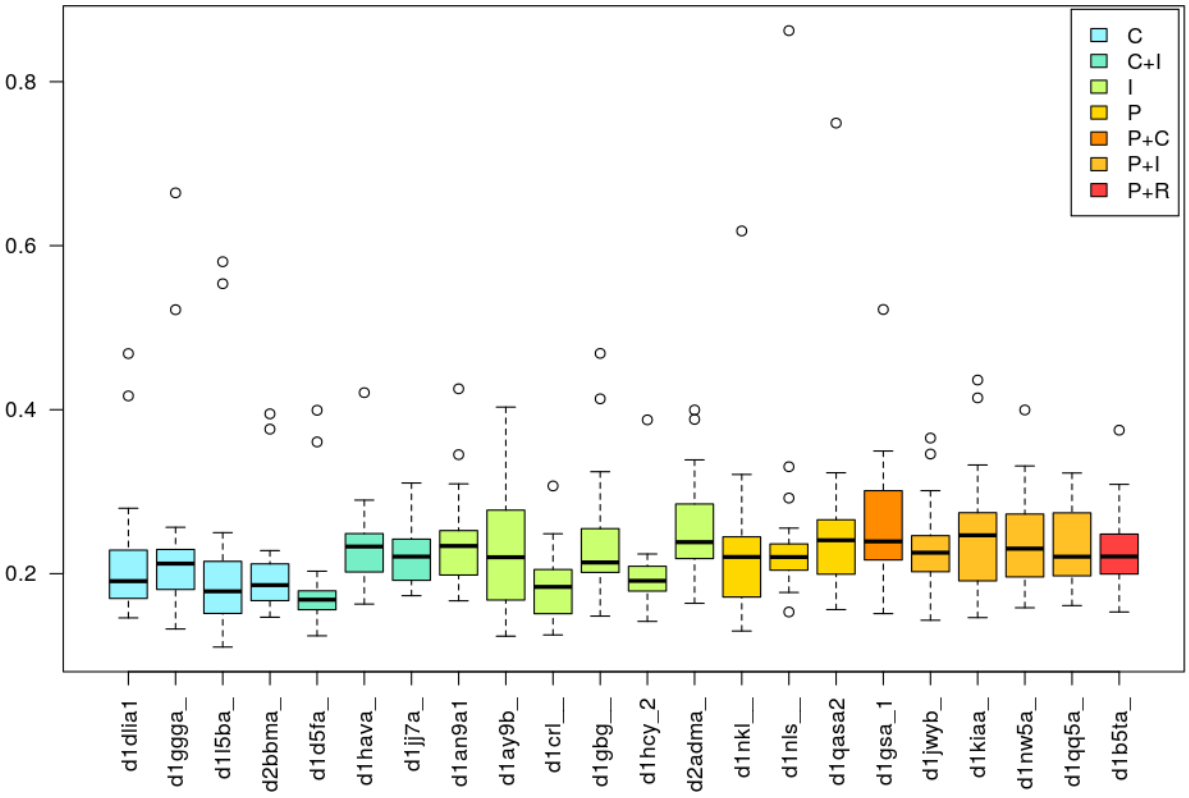


Abbildung 68: RIPC QTMCSS. Die Boxplot-Darstellung der Daten aus dem Kreuzvergleich Abb. 67a. Die richtig-positiven Paare (Tab. 15) sind anhand des QTMCSS als Ausreißer sichtbar.

für jede QS. Während pro Spalte (QS) der Kreuzvergleichmatrix (Abb. 67a) nur 23 MCSs erfasst werden können, enthält jedes Boxplot alle erkannten MCSs pro Domänenpaar. Die ersten 5 Boxplots (Abb. 68, links) gehören zu den QSs der oben genannten Domänenpaare mit großen Konformationsänderungen. Jede der 5 Verteilungen weist jeweils 2 Ausreißer auf, bei denen es sich um MCSs handelt, die die TP-TS möglichst vollständig abdecken. So

entspricht z.B. der größte Ausreißer mit $QTMCSS = 0.468$ der QS d1DLIa1 dem Domänenpaar d1DLIa1/d1MV8a1 aus der Kreuzvergleichmatrix. Der zweite Ausreißer entspricht der zweiten MCS aus dem Vergleich von d1DLIa1 und d1MV8a1 mit $QTMCSS = 0.417$. Obwohl der QTMCSS für viele TP-Paare sehr niedrig ist ($QTMCSS > 0.3$), gibt es signifikante Unterschiede zu der Bewertung der FP-Paare. Das Feld lässt sich mittels z-score

$$z = \frac{s - \mu}{\sigma}$$

standardisieren, mit μ als Mittelwert und σ als Standardabweichung pro QS-Spalte bzw. QS-Boxplot. Abb. 69 zeigt die statistische Signifikanz des QTMCSS. Für diesen Datensatz gilt,

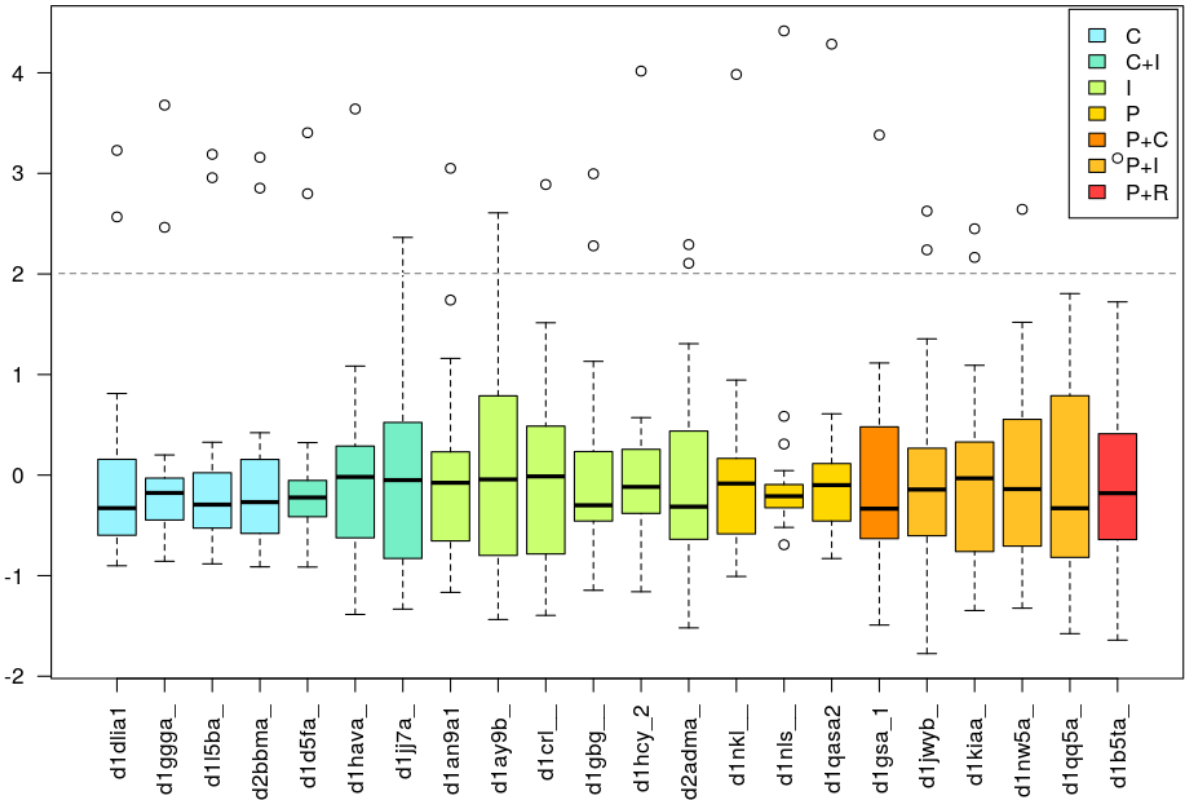


Abbildung 69: RIPC z(QTMCSS). Die Standardisierung des Feldes mit dem z-score offenbart, dass die QTMCSSs der richtig-positiven Paare sich um ≈ 2 (horizontale gestrichelte Linie) Standardabweichungen vom Mittelwert aller betrachteten QTMCSSs unterscheiden.

dass die QTMCSSs der TP-MCSs sich um ≈ 2 (horizontale gestrichelte Linie) Standardabweichungen vom Mittelwert aller betrachteten QTMCSSs unterscheiden. Würde man nach der jeweiligen QS in der gesamten PDB suchen, dann würde sich dieser Wert deutlich nach oben verschieben. Sowohl die signifikante Trennung der TP-MCSs von den FP-MCSs als auch die Clusterung nach den verwandten Faltungsmustern zeigen, dass eine automatisierte Mustererkennung und -klassifizierung anhand des QTMCSS möglich ist.

2.4.3 *Performance*

Die Berechnungen sind auf einem Kern der Xeon X5650-CPU durchgeführt. Abb. 70 demonstriert die Performance von EpitopeMatch für den Vergleich aller 40 Domänenpaare aus dem RIPC-Datensatz. Der Produkt der QS- und der TS-Größen gibt die Anzahl der potentiellen korrespondierenden Residuenpaare pro Domänenpaar an. Der Matching-Modus

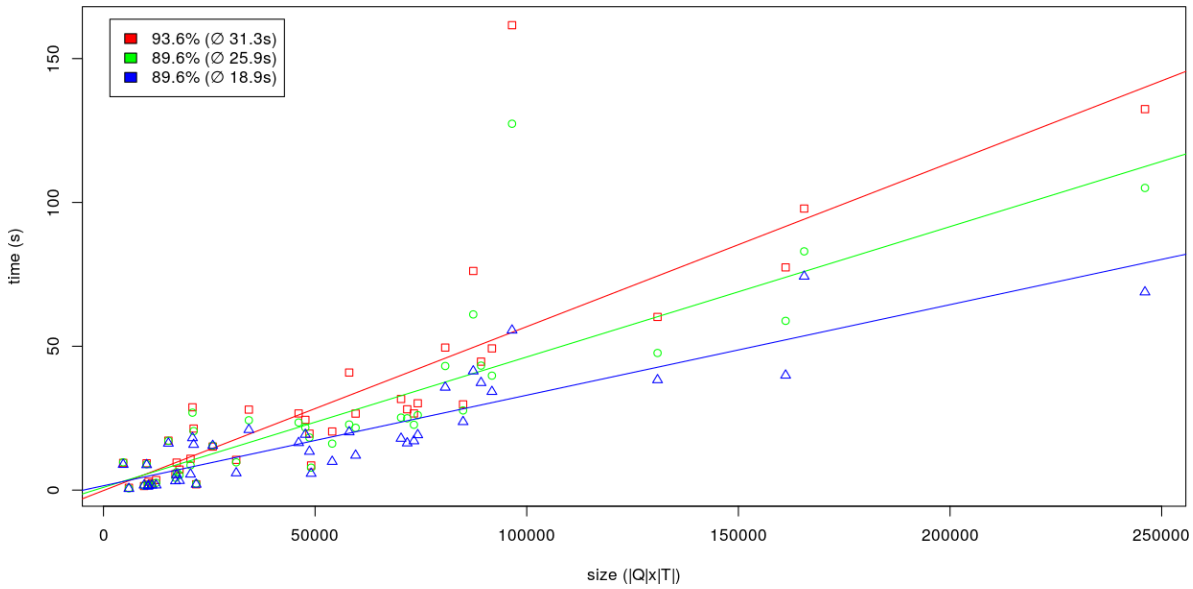


Abbildung 70: RIPC Performance. Die Algorithmus-Performance ist von der Wahl der Templates und von der Anzahl der gespeicherten CSs pro Ordnung k abhängig. Je mehr Informationen die Templates enthalten, desto spezifischer sind die CSs - desto schneller ist der Algorithmus. Je mehr CSs pro Ordnung k gespeichert werden, desto mehr Kind-CSs pro Ordnung $k+1$ werden generiert - desto langsamer der Algorithmus.

ist in allen drei Fällen AllAtoms/Weighted (AA/W). Der Unterschied liegt in der Anzahl der Permutationen pro Ordnung k (Abb. 15), die über den Parameter $BCST_{min}$ getriggert wird. Im Fall der extensiven Suche (rot) ist $BCST_{min} = 3$ (Standardeinstellung), im Fall einer normalen Suche (grün) ist $BCST_{min} = 2$ und im Fall einer schnellen Suche (blau) ist $BCST_{min} = 1$. Die Durchschnittszeit pro Domänenpaar liegt im schnellsten Fall bei 18.9s, wobei mit 89.6% etwas weniger Referenzalignments erkannt werden. Der proportionale Anstieg der Rechenzeit zum Produkt der Größen der QS und TS ist annähernd linear, wobei die Rechenzeit unmittelbar von dem Größenverhältnis der zu vergleichenden Strukturen abhängt. Je kleiner das Verhältnis $min(|QS|,|TS|)/max(|QS|,|TS|)$, desto kürzer die Rechenzeit für Strukturen mit dem gleichen Größenprodukt $|QS| \cdot |TS|$.

2.5 ANWENDUNG

2.5.1 *MTase*

In der ersten publizierten Anwendung [49] fand EpitopeMatch seinen Einsatz in der Kristallografie. Die Autoren klären erstmalig die Struktur der Methyltransferase (MTase) $_{TTC}$TrmN des Thermus thermophilus (3TMA) und seine orthologen Strukturen $_{Pf}$Trm14 des Pyrococcus furiosus (3TLJ, 3TM4, 3TM5) auf. Die letzten drei Strukturen sind als Holo-Strukturen mit den jeweiligen Liganden SAH, SAM und SFG kristallisiert (Abb. 71, cyan), die erste hingegen im Apo-Zustand (Abb. 71, grau). Alle drei Liganden binden in einer sehr ähnlichen Konfiguration der Wasserstoffbrückenbindungen und Van-der-Waals-Interaktionen an die zwischen $_{TTC}$TrmN und $_{Pf}$Trm14 gut konservierte, c-terminale RFM-Domäne (katalytische Rossmann-fold MTase, Abb. 71, links). Mit den Transformationsdaten des auf 3TMA erkannten SAM-Epitops (hier mit 28 Residuen der 3TM4 aus der 5.0$\AA$-Umgebung des SAM) gelang die Positionierung des SAM-Liganden an der entsprechenden Bindungsstelle der 3TMA und somit ein grobes Modell der Holo-Struktur der $_{TTC}$TrmN-MTase. Das Modell

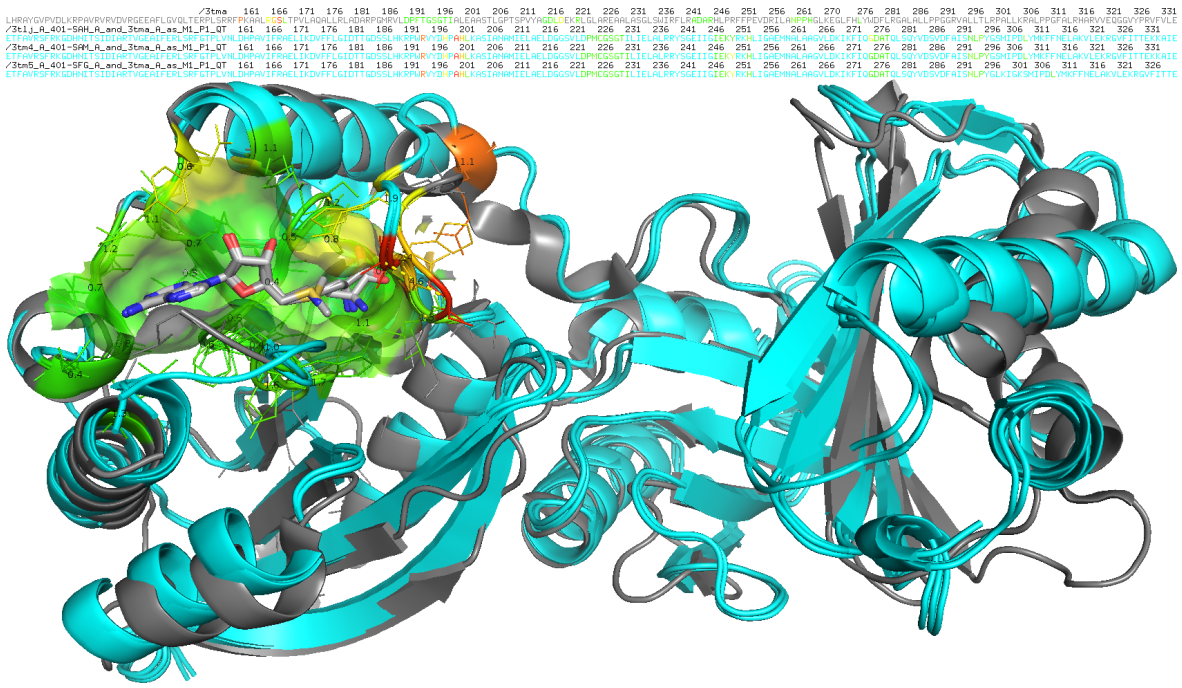


Abbildung 71: MTase. Nachweis einer sehr hohen Ähnlichkeit der Bindestelle auf $_{TTC}$TrmN des Thermus thermophilus (3TMA) zu den Bindestellen der orthologen Strukturen $_{Pf}$Trm14 des Pyrococcus furiosus (3TLJ, 3TM4, 3TM5).

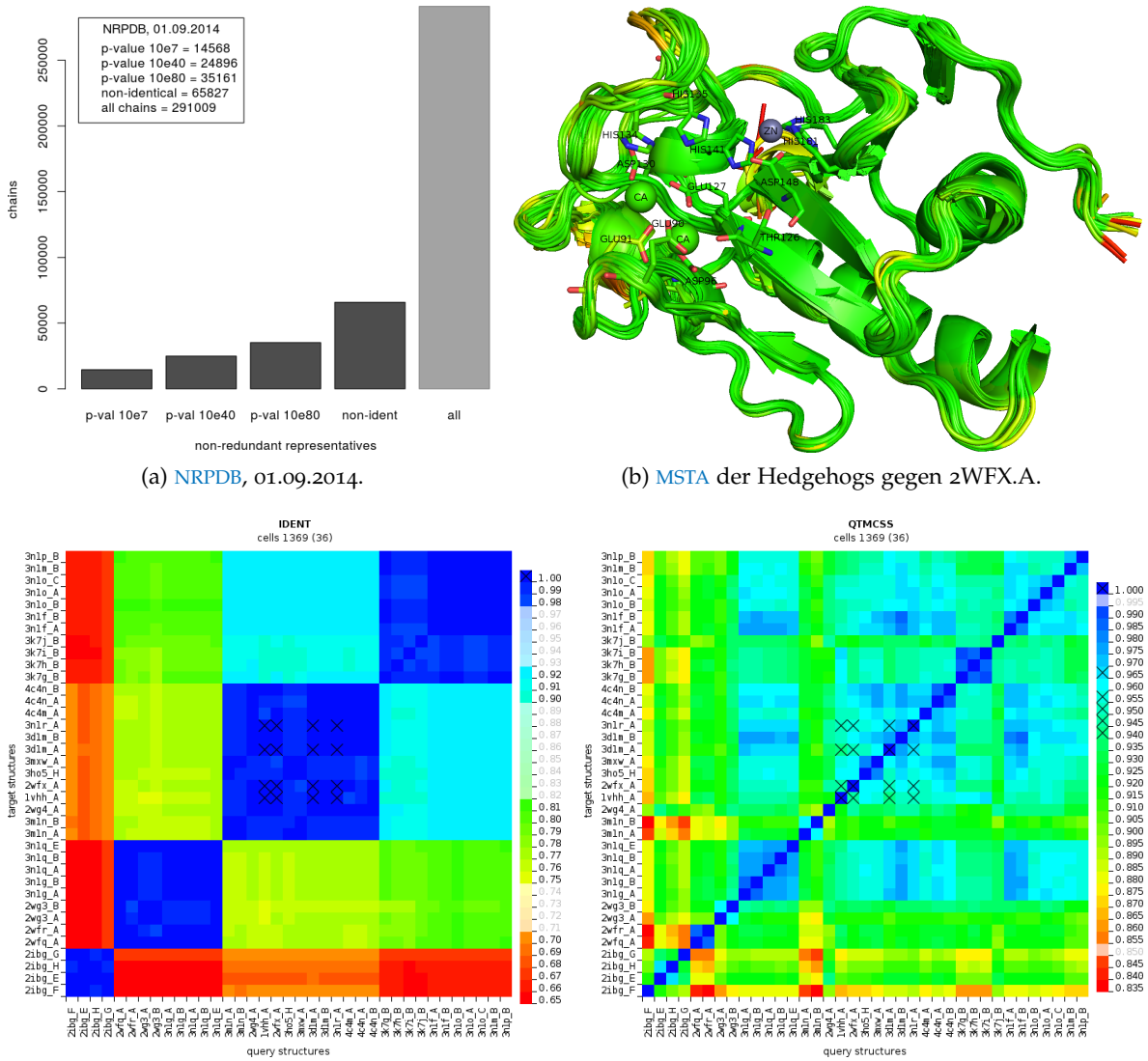
zeigt, dass die entsprechende Bindungsstelle auf $_{TTC}$TrmN mit 28 korrespondierenden Residuen sowohl geometrisch ($RMSD = 1.609\AA$) als auch physiko-chemisch ($IDENT = 0.464$, $SSIM = 0.865$) sehr hohe Ähnlichkeit ($MCSS = 0.906$) zu der SAM-Bindungsstelle auf $_{Pf}$Trm14 aufweist. Darüber hinaus konnte anhand des GMP-Epitops der RsmC-Struktur (3DMH) die Existenz der Bindungsstelle für ein G6-Nukleosid der tRNA nachgewiesen werden. Durch die Erkennung der Bindungsstellen der Holo-Strukturen, und die sich daraus ergebende Transformation der Liganden auf die Apo-Strukturen, leistete EpitopeMatch einen Beitrag zur Aufklärung des Interaktionsmechanismus zwischen der $_{Pf}$Trm14/$_{TTC}$TrmN und der tRNA.

2.5.2 *ShhN*

Im zweiten Anwendungsfall [138] klären die Autoren die Funktion der Sonic Hedgehog, N-terminal domain (ShhN) auf, die vor der Aufklärung als einziger Mitglied der LAS-Enzyme [26] ohne der nachgewiesenen Peptidaseaktivität galt. EpitopeMatch geht in diesem Rahmen der Frage nach, ob abgesehen von den Enzymen weitere Proteine über das zum ShhN ähnliche Zinkzentrum verfügen?

2.5.2.1 *Faltungsmuster*

Die NRPDB (Stand, 01.09.2014) teilt insgesamt 291009 Ketten der PDB (Stand, 01.09.2014) in 14568 Gruppen mit einer niedrigen Redundanz (BLAST p-value 10e-7), in 24896 Gruppen mit einer mittleren Redundanz (BLAST p-value 10e-40), in 35161 Gruppen mit einer hohen Redundanz (BLAST p-value 10e-80) und in 65827 Gruppen mit 100% Sequenzidentität, innerhalb der jeweiligen Gruppe (Abb. 72a). NRPDB-Gruppe 4017 (BLAST p-value 10e-7) enthält 37 Hedgehog-Strukturen. Diese verteilen sich auf: 24 *Desert-*, *Indian-* bzw. *Sonic-*



(a) NRPDB, 01.09.2014.

(b) MSTA der Hedgehogs gegen 2WFX.A.

(c) MSTA der Hedgehogs (Kreuzvergleich), IDENT.

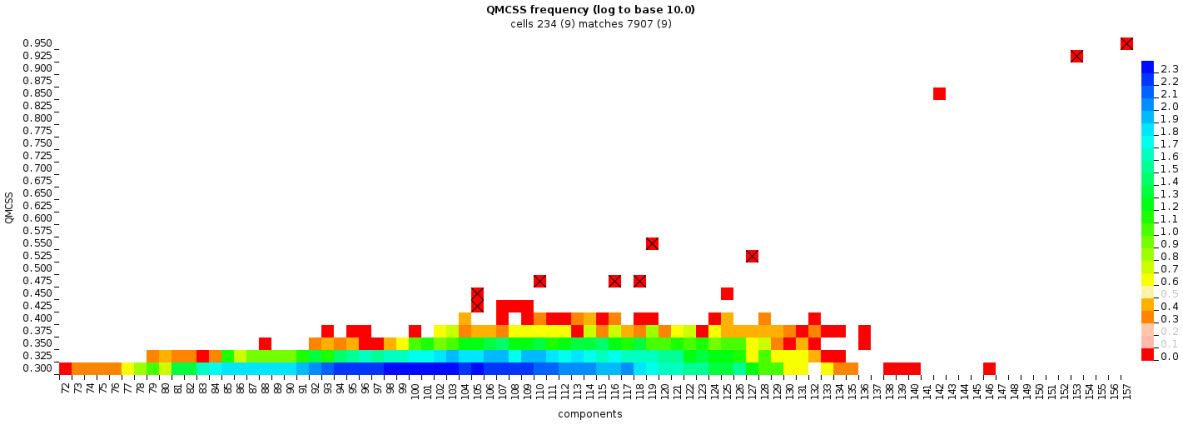
(d) MSTA der Hedgehogs (Kreuzvergleich), QTMCSS.

Abbildung 72: Hedgehogs in der PDB.

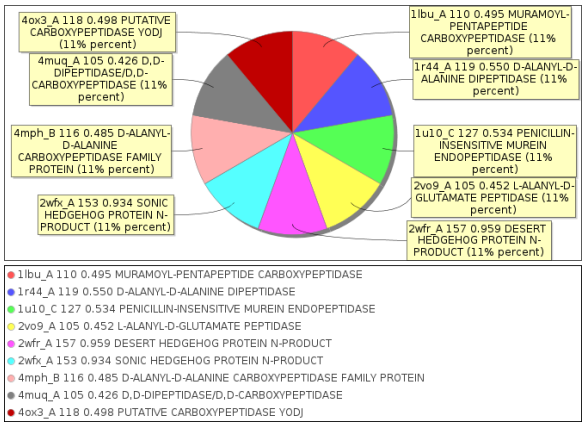
Hedgehogs des *Homo Sapiens*; 9 *Sonic*-Hedgehogs der *Mus Musculus* (Maus); und 4 *Tiggy-Winkle*-Hedgehogs der *Drosophila Melanogaster* (Fruchtfliege). 33 der 37 Hedgehogs verfügen über ein diskontinuierliches H-x-H-Zn^{2+}-Motiv [26], wobei x im Fall der Hedgehogs einem ASP entspricht. Die vierte Koordinationsseite des Zink-Ions füllt ein Wassermolekül. Das H-x-H-Motiv der 4 Hedgehogs der Fruchtfliege ist zu einem H-T-Y mutiert, womit das Zn^{2+}-Motiv verloren gegangen ist. Die Struktur 2WFX.A ist eine ShhN der Maus mit dem höchsten Rang innerhalb der NRPDB-Gruppe 4017 und ist sowohl mit einem Zn^{2+}-Ion als auch mit zwei Ca^{2+}-Ionen (Zustand Ca2) kristallisiert. Als eine Repräsentative der ShhNs innerhalb der Gruppe kann sie als eine Konsensus-Struktur für das MSTA deklariert werden. Abb. 72b demonstriert 37 Matches der 2WFX.A gegen die gesamte NRPDB-Gruppe 4017. 36 Hedgehogs sind mit der 2WFX.A überlagert. 12 Aminosäuren der 2WFX.A (*sticks*) ragen mit mindestens einem Seitenkettenatom (Histidine, Aspartate und Glutamate) oder der Carbonylgruppe (Threonin) in die 5.0Å-Umgebung der Ionen. Im Zustand Ca2 sind insgesamt 18 Hedgehogs (2WFR.A, 2WFX.A, 3D1M.A-B, 3HO5.H, 3MXW.A, 3N1F.A-B, 3N1G.A-B, 3N1M.B, 3N1P.B, 3N1Q.A-B,E, 4C4M.A, 4C4N.A-B). 3N1R.A ist die einzige Struktur mit

nur einem gebundenen Ca^{2+}-Ion (Zustand Ca1). Die restlichen 18 Strukturen sind ohne Ca^{2+}-Ionen (Zustand Ca0) kristallisiert. Das MSTA gegen die 2WFX.A verrät, dass die größte konformationelle Flexibilität im Bereich E131-E138 der 2 konsekutiven Histidine H135 und H136 vorliegt (Abb. 72b). Abb. 72c und Abb. 72d zeigen das vollständige MSTA der Hedgehog-Strukturen. Beide Graphen sind nach IDENT und QTMCSS kombiniert geclustert, sodass die den x- und y-Achsen zugrunde liegenden Dendrogramme in beiden symmetrischen Graphen identisch sind. Verfolgt man die Diagonale in der Abb. 72c von links unten nach rechts oben, so findet man insgesamt 4 Cluster mit: 4 *Tiggy-Winkle*-Hedgehogs der Fruchtfliege; 9 humanen *Desert*-Hedgehogs; 4 humanen *Sonic*-Hedgehogs und 9 *Sonic*-Hedgehogs der Maus; und 11 humanen *Indian*-Hedgehogs. Während der IDENT-Unterschied der *Sonic*- von den *Indian*-Hedgehogs bei $\approx$ 10% liegt, unterscheiden sie sich von den *Desert*- zu $\approx$ 20% und von den *Tiggy-Winkle*-Hedgehogs zu $\approx$ 30%. Zwischen den humanen ShhNs und den ShhNs der Maus liegt der IDENT-Unterschied bei $\approx$ 2%. Die 4 selektierten ShhNs der Maus sind identisch, und repräsentativ für Ca0- (1VHH.A), Ca1- (3N1R.A) und Ca2-Zustand (2WFX.A und 3D1M.A). Auch in der größenabhängigen Bewertung der größten gemeinsamen Substrukturen (Abb. 72d) sind die vier ausgewählten Repräsentativen mit $0.94 < QTMCSS < 0.97$ einander sehr ähnlich.

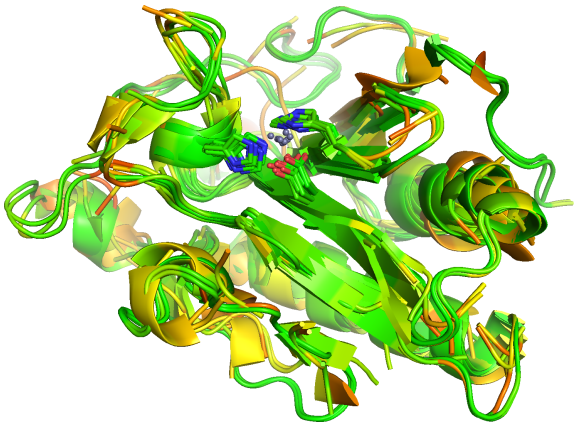
Die 37 Hedgehogs verteilen sich in der NRPDB mit der hohen Redundanz (BLAST p-value



(a) QMCSS-Frequenz der 1VHH.A.



(b) LAS-Repräsentativen.



(c) BCSs der LAS-Repräsentativen.

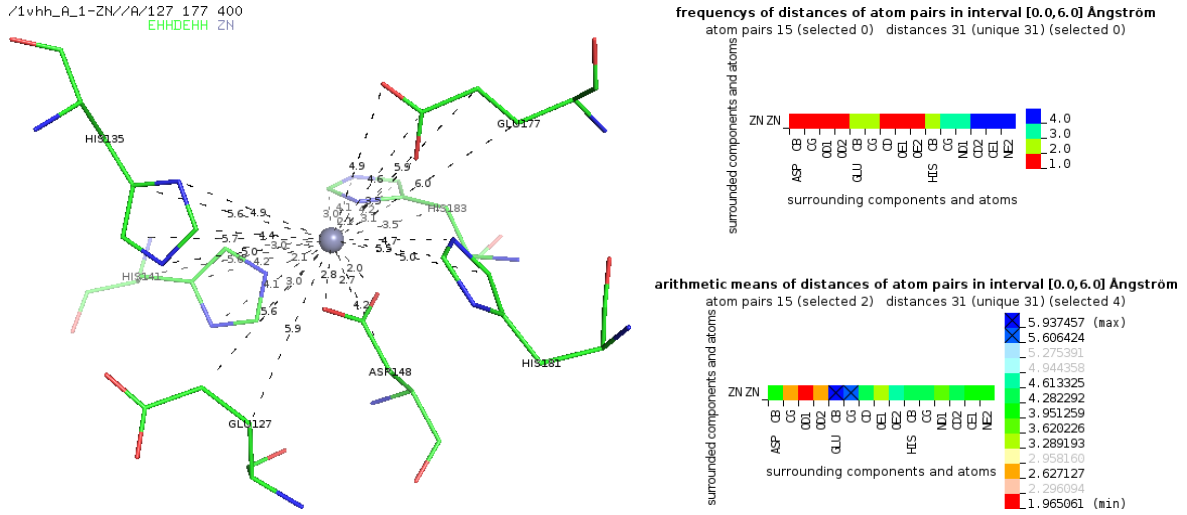
Abbildung 73: LAS-Repräsentativen vs. 1VHH.A.

10e-80) über drei Gruppen: 5544 mit 24 *Sonic*- und *Indian*-Hedgehogs (entspricht der Kombination der beiden größten Cluster in der Abb. 72c); 5545 mit den 9 *Desert*-Hedgehogs; und 24051 mit den 4 *Tiggy-Winkle*-Hedgehogs. Die Rangreihenfolge der ausgewählten Re-

präsentativen innerhalb der Gruppe 5544 ist: 1 2WFX.A; 2 1VHH.A; 4 3N1R.A; und 20 3D1M.A. Abb. 73a stellt die Ergebnisse der Suche nach der 1VHH.A in der NRPDB mit der hohen Redundanz (BLAST p-value 10e-80, 35161 Strukturen), bzw. die QMCSS-Frequenz der größten gemeinsamen Substrukturen zu 1VHH.A. Die Berechnung erfolgte im Modus AllAtoms/AhMwNc7 auf 12 CPU-Kernen zweier Xeon X5650 innerhalb von 6h50m7s. 1VHH.A ähnelt am stärksten der ShhN-Repräsentative 2WFX.A mit $QMCSS = 0.965$, der *Desert*-Hedgehog-Repräsentative 2WFR.A mit $QMCSS = 0.959$ und der *Tiggy-Winkle*-Hedgehog-Repräesentative 2IBG.E mit $QMCSS = 0.857$. Die Letztere ist nicht selektiert, da sie kein Zinkzentrum besitzt. Die Ähnlichkeit zu den 7 weiteren repräsentativen der LAS-Enzyme mit Zinkzentren verteilt sich zwischen $0.426 \leq QMCSS \leq 0.55$. Diese sind ebenfalls selektiert und heben sich als Ausreißer aus der gesamten Verteilung hervor. Abb. 73b zeigt alle LAS-Repräsentativen mit den entsprechenden Bezeichnungen. Alle 7 nicht Hedgehogs sind Peptidasen. Keine Nichtenzyme sind enthalten. Abb. 73c stellt das MSTA der 9 LAS-Repräsentativen gegen die 1VHH.A dar, wobei nur die besten gemeinsamen Substrukturen miteinander überlagert sind. Deutlich erkennbar ist der gemeinsame Kern der Faltungsmuster (grün), der nach CATH die Topologie "Muramoyl-pentapeptide Carboxypeptidase; domain 2" und die Architektur "2-Layer Sandwich" besitzt, und der Klasse "Alpha Beta" zugeordnet ist. Alle 9 LAS-Repräsentativen konservieren das H-x-H-Zn^{2+}-Motiv [26], im Speziellen das HDH-Motiv (*sticks*).

2.5.2.2 *Zinkzentrum*

In die 6.0*Å*-Umgebung des Zinks der 1VHH.A ragen mit ihren Seitenketten insgesamt 7 Aminosäuren der Typen E, D und H, die das vermutete katalytische EHHDEHH-Motiv (E127, H135, H141, D148, E177, H181, H183) darstellen (Abb. 74a). Abgesehen von dem E127



(a) EHHDEHH aus 6.0*Å*-Umgebung.

(b) Frequenz und Distanzen der korresp. Atome

Abbildung 74: Zinkmotiv der 1VHH.A.

(Abb. 74b, selektiert), die im Ca2-Zustand beide Kalziumionen koordiniert und auch im Ca-Apo-Zustand vom Zinkion abgewandt ist, befinden sich alle E-, D- und H-Seitenketten mit mindestens einem Hotspotatom (E.Oε1, E.Oε2, D.Oδ1, D.Oδ2, H.Nδ1, H.Nε2, Abb. 74b) in der 5.0*Å*-Umgebung des Zinkions. Man kann davon ausgehen, dass alle genannten Residuen an dem Mechanismus der enzymatischen Funktion der ShhN maßgeblich beteiligt sind. Im Vergleich zu der Suche nach dem vollständigen ShhN-Faltungsmuster ist die Wahrschein-

lichkeit des häufigeren Vorkommens eines deutlich kleineren Musters des Zinkmotivs höher. Die erste in Betracht gezogene Suchvariante ist faltungsmusterabhängig. Abb. 75a zeigt ein

components and their atoms

0 / 21 0 / 388 21 / 21 168 / 388

21	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
...																					
ASP 1	TYPE	N	CA	C	O	CB	CG	OD1	OD2	OXT	H	H2	HA	HB2	HB3	HD2	HXT				
ASP 1		1	2	3	4	5	6	7	7												
...																					
GLU 1	TYPE	N	CA	C	O	CB	CG	CD	OE1	OE2	OXT	H	H2	HA	HB2	HB3	HG2	HG3	HE2	HXT	
GLU 1		1	2	3	4	5	6	7	7	7											
...																					
HIS 1	TYPE	N	CA	C	O	CB	CG	ND1	CD2	CE1	NE2	OXT	H	H2	HA	HB2	HB3	HD1	HD2	HE1	HE2
HIS 1		1	2	3	4	5	6	7	7	7	7										
...																					
ZN 1	TYPE	ZN																			
ZN 1		7																			

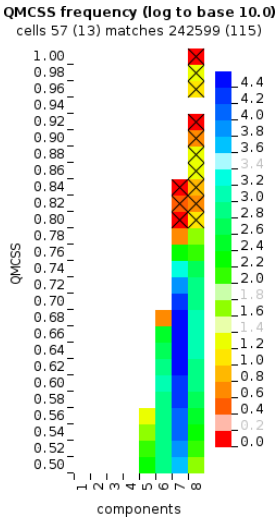
(a) AllAtomsZn-Template.

substitution similarity matrix

21 / 21 / 401 / 441

21	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	ALA 1	ARG 1	ASN 1	ASP 1	CYS 1	GLN 1	GLU 1	GLY 1	HIS 1	ILE 1	LEU 1	LYS 1	MET 1	PHE 1	PRO 1	SER 1	THR 1	TRP 1	TYR 1	VAL 1	ZN 1
ALA 1	1.000	0.381	0.690	0.523	0.881	0.656	0.487	0.882	0.630	0.791	0.817	0.476	0.840	0.766	0.807	0.862	0.830	0.603	0.646	0.839	
ARG 1	0.381	1.000	0.686	0.524	0.428	0.724	0.560	0.426	0.751	0.390	0.415	0.905	0.532	0.539	0.574	0.518	0.551	0.623	0.696	0.364	
ASN 1	0.690	0.686	1.000	0.832	0.742	0.962	0.797	0.736	0.913	0.699	0.725	0.781	0.754	0.680	0.883	0.829	0.861	0.716	0.791	0.674	
ASP 1	0.523	0.524	0.832	1.000	0.590	0.800	0.964	0.568	0.750	0.531	0.557	0.618	0.592	0.517	0.716	0.661	0.693	0.553	0.628	0.506	
CYS 1	0.881	0.428	0.742	0.590	1.000	0.704	0.554	0.764	0.677	0.890	0.916	0.523	0.896	0.866	0.822	0.827	0.866	0.650	0.695	0.917	
GLN 1	0.656	0.724	0.962	0.800	0.704	1.000	0.831	0.701	0.950	0.664	0.690	0.818	0.792	0.718	0.848	0.794	0.826	0.753	0.828	0.639	
GLU 1	0.487	0.560	0.797	0.964	0.554	0.831	1.000	0.532	0.787	0.496	0.522	0.650	0.628	0.554	0.680	0.625	0.657	0.590	0.665	0.471	
GLY 1	0.882	0.426	0.736	0.568	0.764	0.701	0.532	1.000	0.675	0.674	0.700	0.521	0.723	0.648	0.852	0.907	0.875	0.648	0.692	0.721	
HIS 1	0.630	0.751	0.913	0.750	0.677	0.950	0.787	0.675	1.000	0.638	0.664	0.800	0.781	0.737	0.823	0.767	0.800	0.773	0.847	0.613	
ILE 1	0.791	0.390	0.699	0.531	0.890	0.664	0.496	0.674	0.638	1.000	0.974	0.484	0.856	0.848	0.732	0.736	0.776	0.611	0.655	0.953	
LEU 1	0.817	0.415	0.725	0.557	0.916	0.690	0.522	0.700	0.664	0.974	1.000	0.510	0.882	0.874	0.758	0.762	0.802	0.637	0.681	0.949	
LYS 1	0.476	0.905	0.781	0.618	0.523	0.818	0.650	0.521	0.800	0.484	0.510	1.000	0.611	0.536	0.669	0.613	0.646	0.573	0.647	0.459	
MET 1	0.840	0.532	0.754	0.592	0.896	0.792	0.628	0.723	0.781	0.856	0.882	0.611	1.000	0.925	0.781	0.786	0.825	0.754	0.799	0.831	
PHE 1	0.766	0.539	0.680	0.517	0.866	0.718	0.554	0.648	0.737	0.848	0.874	0.536	0.925	1.000	0.707	0.712	0.751	0.762	0.807	0.823	
PRO 1	0.807	0.574	0.883	0.716	0.822	0.848	0.680	0.852	0.823	0.732	0.758	0.669	0.781	0.707	1.000	0.943	0.956	0.743	0.817	0.780	
SER 1	0.862	0.518	0.829	0.661	0.827	0.794	0.625	0.907	0.767	0.736	0.762	0.613	0.786	0.712	0.943	1.000	0.959	0.740	0.785	0.783	
THR 1	0.830	0.551	0.861	0.693	0.866	0.826	0.657	0.875	0.800	0.776	0.802	0.646	0.825	0.751	0.956	0.959	1.000	0.773	0.817	0.813	
TRP 1	0.603	0.623	0.716	0.553	0.650	0.753	0.590	0.648	0.773	0.611	0.637	0.573	0.754	0.762	0.743	0.740	0.773	1.000	0.925	0.586	
TYR 1	0.646	0.696	0.791	0.628	0.695	0.828	0.665	0.692	0.847	0.655	0.681	0.647	0.799	0.807	0.817	0.785	0.817	0.925	1.000	0.630	
VAL 1	0.839	0.364	0.674	0.506	0.917	0.639	0.471	0.721	0.613	0.953	0.949	0.459	0.831	0.823	0.780	0.783	0.813	0.586	0.630	1.000	
ZN 1																					1.000

(b) AhMwNc7Zn-Substitutionsmatrix.



(c) QMCSS

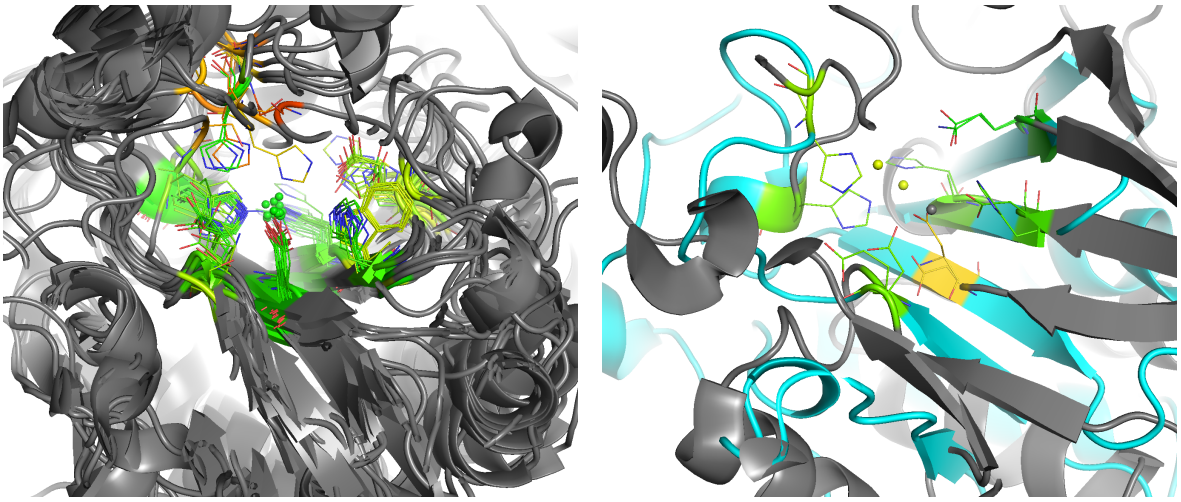
SID_CID	GID 10e80	Rank 10e80	RMSD	IDENT	SSIM	QMCSS	ALIGN EHHDEHHZN	MOL_NAME
2wfr_A	5545	1	0.553	1	1	0.979	EHHDEHHZN	DESERT HEDGEHOG PROTEIN N-PRODUCT
2wfx_A	5544	1	0.646	1	1	0.971	EHHDEHHZN	SONIC HEDGEHOG PROTEIN N-PRODUCT
1lbu_A	20395	1	1.155	0.625	0.874	0.932	GGHDHHHZN	MURAMOYL-PENTAPEPTIDE CARBOXYPEPTIDASE
3csq_D	6888	6	1.509	0.625	0.923	0.907	QSHDHHHZN	MORPHOGENESIS PROTEIN 1
1u10_C	5434	1	1.646	0.75	0.969	0.906	DHHDHHHZN	PENICILLIN-INSENSITIVE MUREIN ENDOPEPTIDASE
1qwy_A	22494	1	1.618	0.625	0.921	0.9	QDHDHHHZN	PEPTIDOGLYCAN HYDROLASE
4lxc_A	5143	1	1.659	0.625	0.889	0.892	YLHDHHHZN	LYSOSTAPHIN
4muq_A	7384	1	1.606	0.625	0.879	0.89	SAHDEWHZN	D,D-DIPEPTIDASE/D,D-CARBOXYPEPTIDASE
1r44_A	22568	1	1.676	0.625	0.921	0.883	DAHDEWHZN	D-ALANYL-D-ALANINE DIPEPTIDASE
4mur_A	33720	3	1.648	0.625	0.879	0.883	SAHDEWHZN	D,D-DIPEPTIDASE/D,D-CARBOXYPEPTIDASE
2b44_A	22498	1	2.068	0.625	0.941	0.879	QNHDHHHZN	GLYCYL-GLYCINE ENDOPEPTIDASE LYTM
4ox3_A	33719	1	1.761	0.625	0.879	0.873	SAHDEWHZN	PUTATIVE CARBOXYPEPTIDASE YODJ
2vo9_A	29849	1	1.771	0.5	0.906	0.872	QAHDDPHZN	L-ALANYL-D-GLUTAMATE PEPTIDASE
4d0y_A	7383	1	1.652	0.625	0.879	0.872	SAHDEWHZN	DACB
2gu1_A	22497	1	2.099	0.625	0.878	0.868	SVHDHHHZN	ZINC PEPTIDASE
2ptz_A	3032	32	1.751	0.25	0.816	0.864	DGGDQDAZN	ENOLASE
4mph_B	5089	1	1.791	0.625	0.879	0.864	SAHDEWHZN	D-ALANYL-D-ALANINE CARBOXYPEPTIDASE FAMILY PROTEIN
3it5_A	32081	1	1.805	0.625	0.907	0.856	NGHDHHHZN	PROTEASE LASA
2ibg_E	24051	1	0.623	0.571	0.859	0.843	EHHTVHY	PROTEIN HEDGEHOG
3vpb_B	23618	1	1.802	0.375	0.843	0.827	ESNDRDRZN	PUTATIVE ACETYLORNITHINE DEACETYLASE
2b13_A	22499	1	1.077	0.714	0.945	0.823	Q HDHHHZN	GLYCYL-GLYCINE ENDOPEPTIDASE LYTM
2bnm_A	5983	2	1.768	0.25	0.695	0.817	MHVLVLAZN	EPOXIDASE
3fid_A	31537	1	1.903	0.375	0.823	0.815	TLKDHHGZN	PUTATIVE OUTER MEMBRANE PROTEIN (LPXR)
2q1z_D	6477	1	1.818	0.375	0.769	0.814	EHATLCAZN	ANTI-SIGMA FACTOR CHRR, TRANSCRIPTIONAL ACTIVATOR CHRR
2vad_A	1726	20	1.862	0.25	0.719	0.811	EYLVVAAZN	RED FLUORESCENT PROTEIN
3rcq_A	33086	1	2.127	0.25	0.86	0.806	QEKDDRRZN	ASPARTYL/ASPARAGINYL BETA-HYDROXYLASE
2hsi_B	27798	1	2.537	0.625	0.904	0.805	SPHDHHHZN	PUTATIVE PEPTIDASE M23
2anu_E	25856	4	2.022	0.5	0.856	0.801	HHHEAHVZN	HYPOTHETICAL PROTEIN TM0559
2hpm_A	6270	19	2.003	0.5	0.885	0.8	HHHEGHTZN	DNA POLYMERASE III ALPHA SUBUNIT

(d) 29 Repräsentativen (BLAST p-value 10e-80).

Abbildung 75: AllAtomsZn/AhMwNc7Zn-Modus (faltungsmusterabhängig).

Auszug aus dem AllAtomsZn-Template, das ein modifiziertes AllAtoms-Template (Abb. 5) darstellt, in dem die Korrespondenzen zwischen den Atomen der Residuen E, D und H, und dem Zinkion, auf 7 Ebenen verteilt sind. Die Hotspotatome der 3 Aminosäuren sind als geometrische Zentren zusammengefasst und korrespondieren mit dem Zinkion in der

siebten Distanzmatrix. Abb. 75c zeigt die von der Größe des Epitops abhängige Frequenz der größten gemeinsamen Substrukturen (QMCSS) in der gesamten PDB (291009 Ketten). Die Suche nach dem Zinkmotiv (Abb. 74a) nahm im Modus AllAtomsZn/AhMwNc7Zn auf 12 CPU-Kernen zweier Xeon X5650 1h13m58s in Anspruch. Im Bereich $0.8 < QMCSS \leq 1.0$ befinden sich 115 Ausreißer-MCSs, die sich auf 29 NRPDB-Gruppen (BLAST p-value 10e-80) verteilen. Jede Gruppe wird durch eine Struktur mit dem höchsten Rang vertreten (Abb. 75d). Würde man die Suche nur auf den Repräsentativen der NRPDB-Gruppen (BLAST p-value 10e-80) durchführen, so würden alle MCSs der Strukturen mit dem $Rank > 1$ (7 von 29) unerkannt bleiben. Aus diesem Grund empfiehlt sich für die Suche nach Epitopen immer eine Suche in der vollständigen PDB. Die Liste in der Abb. 75d ist nach dem QMCSS absteigend sortiert. Der QMCSS vereint die Geometrie (RMSD), die Substitutionsähnlichkeit (SSIM) und die Größe des Alignments (ALIGN). Die Spalte ALIGN stellt ein MSTA gegen das diskontinuierliche Konsensus-Epitop der 1VHH.A dar. Das HDH-Zinkmotiv des Epitops ist rot hervorgehoben und wird mit Ausnahme von 2PTZ.A (Enolase) von allen Strukturen bis zu der 2IBG.E (*Tiggy-Winkle*-Hedgehog) konserviert. Abb. 76a zeigt die



(a) Zinkmotive der Peptidasen.

(b) Zinkmotive der 1VHH.A & 2PTZ.A.

Abbildung 76: Zinkmotive (faltungsmusterabhängig).

Überlagerung der ersten 17 LAS-Enzyme mit dem 1VHH.A-Zinkmotiv. Die Koordination der Zinkionen (grün) ist sehr ähnlich. Die physiko-chemische Ähnlichkeit des Zinkmotivs schwankt im Bereich $0.874 \leq SSIM \leq 1.0$ (bei $0.5 \leq IDENT \leq 1.0$), die geometrische im Bereich $0.553\mathring{A} \leq RMSD \leq 2.099\mathring{A}$. Im Fall der Enolase (Abb. 76b) mit $SSIM = 0.816$ ($IDENT = 0.25$) und $RMSD = 1.751\mathring{A}$ verändert sich die Koordination des Zinkions (gelb). Darüber hinaus koordiniert das Zinkmotiv der Enolase ein zweites Zinkion (grau). Mit der fallenden physiko-chemischen Ähnlichkeit bzw. steigenden Mutationsrate findet man immer mehr MCSs, die der faltungsmusterabhängigen Geometrie des 1VHH.A-Zinkmotivs eher zufallsbedingt und dem Mechanismus der enzymatischen Funktion der ShhN immer weniger entsprechen. Der Trend bleibt jedoch der gleiche - 14 besten MCSs (bis 2IBG.E, Abb. 75d), ausgenommen die Hedgehogs und die Enolase, sind Peptidasen. Keine Nichtenzyme sind enthalten.

Die zweite Suchvariante ist faltungsmusterunabhängig. Die geometrischen Zentren der Hotspotatome des EHHDEHH-Motivs und das Zinkion sind in einer einzigen Distanzmatrix zusammengefasst (Abb. 77a). E und D sind vom Typ 1, H vom Typ 2 und Zn vom Typ 3. Diese Typisierung der chemischen Komponenten resultiert in einer spezifischen Substi-

components and their atoms

0 / 4 0 / 57 4 / 4 11 / 57

4	1	2	3	4	5	6	7	8	9	10	11
ASP 1	TYPE	N	CA	C	O	CB	CG	OD1	OD2	OXT	H
ASP 1	1							1	1		
GLU 1	TYPE	N	CA	C	O	CB	CG	CD	OE1	OE2	OXT
GLU 1	1								1	1	
HIS 1	TYPE	N	CA	C	O	CB	CG	ND1	CD2	CE1	NE2
HIS 1	2							1			1
ZN 1	TYPE	ZN									
ZN 1	3	1									

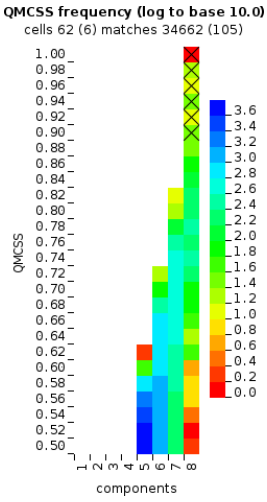
(a) HotspotsZn-Template.

substitution similarity matrix

4 / 4 / 6 / 16

4	1	2	3	4
	ASP 1	GLU 1	HIS 1	ZN 1
ASP 1	1.000	1.000	0.000	0.000
GLU 1	1.000	1.000	0.000	0.000
HIS 1	0.000	0.000	1.000	0.000
ZN 1	0.000	0.000	0.000	1.000

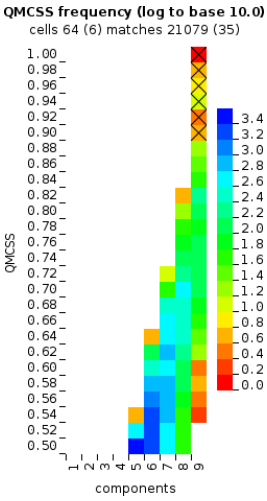
(b) SpecificZn-Substitutionsmatrix.



(c) QMCSS

SID_CID	GID 10e80	Rank 10e80	RMSD	IDENT	SSIM	QMCSS	ALIGN EHHDEHHZN	MOL_NAME
2wfq_A	5545	2	0.92	1	1	0.975	EHHDEHHZN	DESERT HEDGEHOG PROTEIN N-PRODUCT
2i7v_A	26754	1	0.969	0.875	1	0.956	EHHDDHHZN	CLEAVAGE AND POLYADENYLATION SPECIFICITY FACTOR 73 KDA SUBUNIT
3zwf_A	25148	1	0.923	0.875	1	0.956	EHHDDHHZN	ZINC PHOSPHODIESTERASE ELAC PROTEIN 1
2wfx_A	5544	1	1.147	1	1	0.953	EHHDEHHZN	SONIC HEDGEHOG PROTEIN N-PRODUCT
3iek_A	6083	1	1.071	0.875	1	0.951	EHHDDHHZN	RIBONUCLEASE TTHA0252
1y44_A	5661	5	0.928	0.875	1	0.95	EHHDDHHZN	RIBONUCLEASE Z
2e7y_A	24721	1	0.953	0.875	1	0.948	EHHDDHHZN	TRNASE Z
1zkp_C	25555	4	0.912	0.875	1	0.947	EHHDDHHZN	HYPOTHETICAL PROTEIN BA1088
3zdk_A	30341	1	1.098	0.75	1	0.945	DHHDDHHZN	5' EXONUCLEASE APOLLO
2az4_B	25936	1	0.966	0.875	1	0.942	EHHDDHHZN	HYPOTHETICAL PROTEIN EF2904
3ztv_A	18491	1	1.248	0.75	1	0.941	DHHDDHHZN	NAD NUCLEOTIDASE
3zq4_D	5918	8	1.09	0.75	1	0.933	DHHDDHHZN	RIBONUCLEASE J 1
2anu_B	25856	2	1.494	0.625	1	0.923	DHHEDHHZN	HYPOTHETICAL PROTEIN TM0559
4b87_A	30342	1	1.577	0.75	1	0.919	DHHDDHHZN	DNA CROSS-LINK REPAIR 1A PROTEIN
4le6_D	4993	6	1.524	0.75	1	0.914	DHHDDHHZN	ORGANOPHOSPHORUS HYDROLASE
3ayv_D	30688	1	1.589	0.875	1	0.913	EHHEEHHZN	PUTATIVE UNCHARACTERIZED PROTEIN TTHB071
3e38_B	31290	2	1.404	0.625	1	0.911	DHHEDHHZN	TWO-DOMAIN PROTEIN CONTAINING PREDICTED PHP-LIKE METAL- DEPENDENT PHOSPHOESTERASE
2z1a_A	3883	8	1.736	0.75	1	0.91	DHHDDHHZN	5'-NUCLEOTIDASE
4gc3_A	30471	1	1.476	0.625	1	0.906	DHHEDHHZN	L-HISTIDINOL PHOSPHATE PHOSPHATASE
4jom_A	6270	6	1.54	0.75	1	0.904	DHHDDHHZN	DNA POLYMERASE III SUBUNIT ALPHA
4hno_A	5136	3	1.962	0.75	1	0.903	DHHEEHHZN	PROBABLE ENDONUCLEASE 4
2ycb_B	6084	1	1.455	0.875	1	0.902	EHHDDHHZN	CLEAVAGE AND POLYADENYLATION SPECIFICITY FACTOR
4cog_A	5172	1	1.629	0.875	1	0.901	DHHDEHHZN	KYNURENINE FORMAMIDASE

(d) 23 Repräsentativen (BLAST p-value 10e-80)



(e) QMCSS

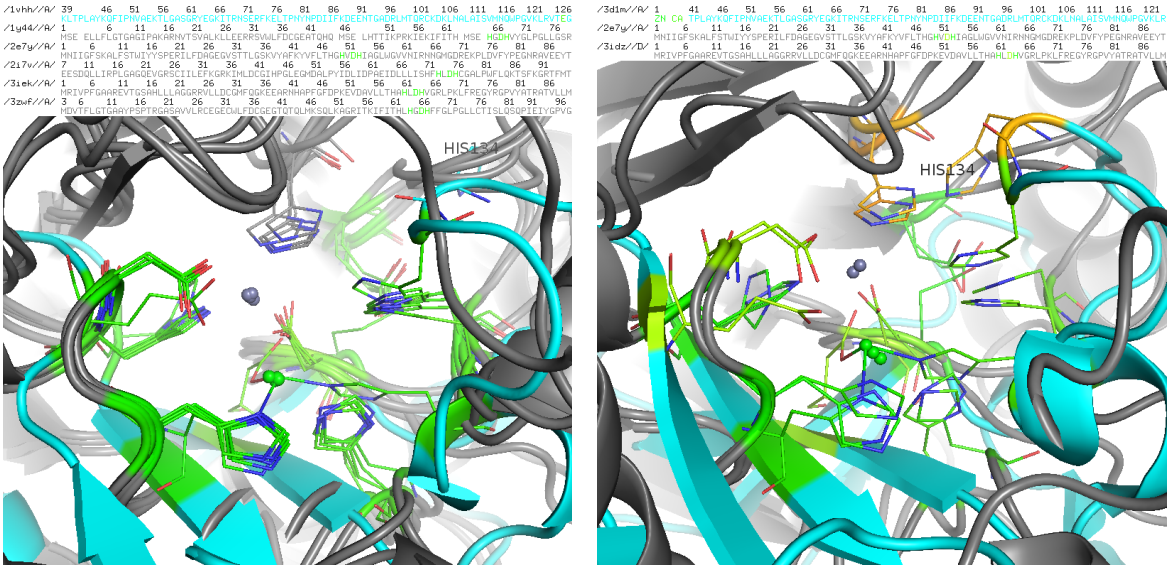
SID_CID	GID 10e80	Rank 10e80	RMSD	IDENT	SSIM	QMCSS	ALIGN EHHHDEHHZN	MOL_NAME
2wfx_A	5544	1	1.358	1	1	0.953	EHHHDEHHZN	SONIC HEDGEHOG PROTEIN N-PRODUCT
2wfr_A	5545	1	1.714	1	1	0.927	EHHHDEHHZN	DESERT HEDGEHOG PROTEIN N-PRODUCT
2e7y_A	24721	1	1.657	0.889	1	0.924	EHHHDDHHZN	TRNASE Z
3idz_D	6083	16	1.696	0.889	1	0.903	EHHHDDHHZN	RIBONUCLEASE TTHA0252
1zkp_B	25555	3	1.879	0.889	1	0.895	EHHHDDHHZN	HYPOTHETICAL PROTEIN BA1088
3zwf_A	25148	1	1.907	0.889	1	0.889	EHHHDDHHZN	ZINC PHOSPHODIESTERASE ELAC PROTEIN 1
3bk1_A	5918	3	1.826	0.778	1	0.887	DHHHDDHHZN	METAL DEPENDENT HYDROLASE
3ayv_D	30688	1	1.955	0.778	1	0.882	EHHHEDHHZN	PUTATIVE UNCHARACTERIZED PROTEIN TTHB071
2fk6_A	5661	1	1.852	0.889	1	0.875	EHHHDDHHZN	RIBONUCLEASE Z
2w9m_B	29951	1	2.444	0.667	1	0.86	DHHHEDHHZN	POLYMERASE X
4jom_A	6270	6	1.896	0.778	1	0.859	DHHHDDHHZN	DNA POLYMERASE III SUBUNIT ALPHA
2cfz_A	6028	2	2.176	0.778	1	0.856	EHHHEDHHZN	SDS HYDROLASE SDSA1
2yxo_B	30470	1	2.158	0.667	1	0.856	DHHHEDHHZN	HISTIDINOL PHOSPHATASE
4jdg_A	1146	3	2.052	0.778	1	0.853	DHHHDDHHZN	NUCLEASE
4dsy_B	4394	1	1.951	0.778	1	0.853	DHHHDDHHZN	PURPLE ACID PHOSPHATASE
4gyf_A	30471	2	2.434	0.667	1	0.853	DHHHEDHHZN	HISTIDINOL-PHOSPHATASE
1p9e_B	4993	2	1.973	0.778	1	0.852	DHHHDDHHZN	METHYL PARATHION HYDROLASE
3mk1_A	1132	3	1.915	0.889	1	0.849	EHHHDDHHZN	ALKALINE PHOSPHATASE, PLACENTAL TYPE
2anu_A	25856	1	2.523	0.667	1	0.846	DHHHEDHHZN	HYPOTHETICAL PROTEIN TM0559
1pb0_B	21915	2	2.012	0.667	1	0.844	DHHHEDHHZN	HYPOTHETICAL PROTEIN YCDX
2x7v_A	5136	1	2.281	0.778	1	0.843	DHHHEEHHZN	PROBABLE ENDONUCLEASE 4
2bib_A	3721	3	2.201	0.778	1	0.842	DHHHDDHHZN	TEICHOIC ACID PHOSPHORYLCHOLINE ESTERASE/ CHOLINE BINDING PROTEIN
1qtw_A	5135	1	2.284	0.778	1	0.841	DHHHEEHHZN	ENDONUCLEASE IV
4ajd_A	3267	15	2.117	0.889	1	0.84	EHHHDDHHZN	CAMP AND CAMP-INHIBITED CGMP 3', 5'-CYCLIC PHOSPHODIESTERASE 10A, PDE10

(f) 24 Repräsentativen (BLAST p-value 10e-80).

Abbildung 77: HotspotsZn/SpecificZn-Modus (faltungsmusterunabhängig).

tutionsmatrix (Abb. 77b), in der die Mutation zwischen E und D zugelassen ist. Die auf Hotspots reduzierte Geometrie und die verringerte Anzahl der möglichen Substitutionen resultieren in einer Rechenzeit von lediglich 6m33s im Modus HotspotsZn/SpecificZn (in der gesamten PDB) auf 12 CPU-Kernen zweier Xeon X5650. Die besten 105 MCSs aus dem Bereich $0.9 < QMCSS \leq 1.0$ enthalten, abgesehen von den Hedgehogs, keine LAS-Enzyme (Abb. 77d). Diese verteilen sich aufgrund der zahlreichen Mutationen (Abb. 75d, ALIGN) im verrauschten Bereich der QMCSS-Verteilung (Abb. 77c). Die 105 MCSs verteilen sich auf 23 NRPDB-Gruppen (BLAST p-value 10e-80). Auffällig ist, dass alle Strukturen (außer 4B87.A und der Hedgehogs 2WFQ.A und 2WFX.A) über sehr ähnliche Hotspotkonformationen

verfügen, die an einer Koordination von mindestens 2 Zinkionen beteiligt sind. Allerdings beteiligt sich an der Koordination des zweiten Zinkions mindestens eine weitere Seitenkette. Abb. 78a zeigt das MSTA der 5 besten MCSs gegen die 1VHH.A (Cao-Zustand). Die



(a) Motive mit 2 Zinkionen (Cao-Zustand).

(b) Motive mit 2 Zinkionen (Ca2-Zustand).

Abbildung 78: Zinkmotive (faltungsmusterunabhängig).

zweiten Zinkionen (grau) werden zusätzlich von einem Histidin der jeweiligen Struktur koordiniert. Auch 1VHH.A besitzt ein Histidin (HIS134) in der Nähe, allerdings ist seine Seitenkette im Cao-Zustand vom katalytischen Zentrum abgewandt. Im Ca2-Zustand (3D1M.A) ist HIS134 dem katalytischen Zentrum zugewandt. Um zu testen, ob das Zinkmotiv der ShhN dem möglichen Bindungsmotiv mit 2 Zinkionen entspricht, ist das HIS134 in das Zinkmotiv aufgenommen worden, wobei das so entstandene EHHHDEHH-Motiv der 3D1M.A entnommen worden ist. Die Suche nach dem erweiterten Zinkmotiv zeigt eine deutlich geringere Übereinstimmung mit den Motiven, die zwei Zinkionen koordinieren. Im Bereich $0.9 < QMCSS \leq 1.0$ (Abb. 77e) befinden sich 33 Hedgehogs (keine *Tiggy-Winkle*-Hedgehogs), die "TRNASE Z" (2E7Y.A) und "RIBONUCLEASE TTHA0252" (3IDZ.D). Bei der näheren Betrachtung der beiden letzten Ausreißer (Abb. 78b) stimmt die Position des H134 (orange) trotz der richtigen Ausrichtung nicht mit der Position der Histidine überein, die die zweiten Zinkionen (grau) koordinieren. Der Trend der QMCSS-Verteilung und die abweichende Position des H134, sprechen gegen seine mögliche Beteiligung an der Koordination eines zweiten Zinkions im katalytischen Zentrum der ShhN. Im Gegensatz zu der faltungsmusterabhängigen Ähnlichkeit des Motivs zu den Peptidasen, ähnelt die faltungsmusterunabhängige Konformation der Hotspots des Motivs den Phosphodiesterasen. Auch hier ist keine Ähnlichkeit zu den Nichtenzymen zu verzeichnen.

2.6 DISKUSSION

Die in den letzten vier Abschnitten erklärte und demonstrierte Kernfunktionalität von EpitopeMatch bietet als eine Basisoperation des Strukturmodellierungsprozesses eine umfangreiche Umsetzung der wichtigsten Facetten des SCA:

- Abs. 2.2 - eine heuristische Lösung des klassischen Problems der computergestützten Geometrie - das Vergleichen von Punktwolken im Raum. Die templatebasierte Defi-

nition der geometrischen Objekte erstreckt sich von den Korrespondenzen einzelner Atome (z.B. Cα) über beliebige Atomkombinationen der korrespondierenden Residuen bis hin zu Vertexnormalen der triangulierten Oberflächen. In Kombination mit der physiko-chemischen Gewichtung der korrespondierenden Objekte durch ebenfalls frei definierbaren Substitutionsmatrizen entsteht eine Reihe von Deskriptoren und die eigentliche OF (CSS, Gl. 40), die durch die Implikation der Substrukturinformation die Ähnlichkeit zweier Strukturen signifikant bewertet. Ihre Eignung für das Paarweises-MSTA ist die Grundlage für das Data-Mining (z.B. Clusterung) und somit für die Erkennung der Kreuzreaktivitäten.

- Abs. 2.3 - von der Syntheserichtung unabhängiges, diskontinuierliches Matchen von Biopolymeren, ihren Substrukturen (z.B. Epitope) und einzelnen molekularen Verbindungen (z.B. ATP) unter der Berücksichtigung ihrer Flexibilität (Induced-Fit). Die Hilfswerkzeuge zum Filtern und Ausschneiden der Substrukturen auf der Grundlage der bekannten Komplexstrukturen der PDB erlauben eine flexible Definition von i.a. diskontinuierlichen Epitopen. Ein Satz von vordefinierten Templates und Substitutionsmatrizen gestattet eine schnelle Suche nach den faltungsmusterabhängigen und -unabhängigen Homologien. Es besteht die Möglichkeit der Definition der Epitope in Form von Hotspots und Vertexnormalen (MSMS, [147]). Dem Problem der Multimodalität kann mit dem Pseudoepitop-Ansatz begegnet werden, in dem eine Kombination aus unterschiedlichen Epitopen des gleichen Liganden ein Konsensus-Epitop darstellt und die Erkennung des Epitops auf nicht homologen Apo-Strukturen möglich macht. Die tabellarische, hierarchische Datenerfassung mit den zusammengeführten Daten aus NRPDB, GO, und Taxonomie ermöglicht durch die umfangreichen Filterungsoptionen eine zielgerichtete Ergebnisdarstellung und -analyse in Form von unterschiedlichen Graphen.

- Abs. 2.4 - eine Erweiterung und Verbesserung im Feld der SCA-Methoden. Ungeachtet der Repetitionen, Insertionen/Deletionen, zirkulären Permutationen und beträchtlichen Konformationsänderungen erkennt EpitopeMatch im Vergleich zu den als "State of the Art" geltenden Algorithmen signifikant mehr Ähnlichkeitskerne der entfernt homologen und nicht homologen Strukturen. Gerade die Erkennung solcher Ähnlichkeitskerne führt zu der Identifizierung der evolutionär konservierten Residuen und somit zu einem besseren Verständnis der Wechselwirkungsmechanismen. Tab. 16 stellt eine pessimistische Schätzung der Performance für die möglichen Matching-Szenarien auf einer konventionellen Workstation (Lenovo ThinkStation C20 4263-72G).

Query/Target	Epitop (25)	Struktur (250)	EDB (10^5)	PDB ($2.7 \cdot 10^5$)	NRPDB($1.5 \cdot 10^4$)
Epitop	0.5s	1.5s	1.2h	9h	1h
Struktur	1.5s	15s	3.5h	4d	5h
EDB	1.2h	3.5h	6.5y	107y	6y
PDB	9h	4d	107y	1445y	160y
NRPDB	1h	5h	6y	160y	4.5y

Tabelle 16: Geschätzte Performance der Matching-Szenarien auf 12 CPU-Kernen zweier Xeon X5650. Mittlere Größe eines Epitops bzw. einer Struktur sind 25 bzw. 250 Residuen. Angenommene Größe einer EDB ist 10^5 Epitope. Etwa $0.2 \cdot 10^5$ der $2.9 \cdot 10^5$ Strukturen der PDB enthalten keine Aminosäuren. Grün: in dieser Arbeit gerechneten Szenarien. Gelb bzw. rot: aufgrund der fehlenden EDB bzw. derzeitigen Performance undurchführbaren Szenarien.

Die Schätzung der Größe einer EDB und der mittleren Größe eines Epitops ist schwierig, da die Definition eines Epitops unmittelbar mit der Definition der entsprechenden Bindungsmechanismen zusammenhängt, die in den meisten Fällen nicht klar definiert sind. Die derzeitige Performance von EpitopeMatch ist ausreichend für die effiziente Suche nach einem Epitop oder einer Struktur in der gesamten PDB und somit für eine gezielte Untersuchung eines bestimmten Epitops oder einer Struktur. Aber auch Teilmengen aus der hypothetischen EDB, die mit EpitopeMatch erzeugt werden können, und Teilmengen aus der PDB, können im Sinne von MSTA effizient kreuzvalidiert werden.

- Abs. 2.5 - eine Hilfestellung bei der Aufklärung der Bindungsmechanismen. Anhand der qualitativen und quantitativen Untersuchung der Holo- und Apo-Epitope lassen sich ihre starren und flexiblen Anteile differenzieren. Diese Erkenntnis ist richtungweisend im Bezug auf die Aufklärung des Bindungsmechanismus und kann die präzise Aufklärung beschleunigen. Die Deskriptoren (Abs. 2.2.6.1) bzw. ihre Kombination sind für die Trennung der richtig-positiven Ergebnisse von den falsch-positiven Ergebnissen außerhalb der Zwielichtzone sehr gut geeignet. Darüber hinaus, ist die hypothetische Zuordnung der Proteinfunktion möglich, was eine Hilfe für die gezielte Durchführung der Experimente im Labor sein kann.

Existiert eine Target-Struktur, experimentell bestimmt oder modelliert, und ein Query-Epitop, mit dem die Target-Struktur ausgestattet werden soll, so kann mithilfe von EpitopeMatch im faltungsmusterabhängigen oder -unabhängigen Matching-Modus geprüft werden, ob die Target-Struktur über eine zum Query-Epitop ähnliche, diskontinuierliche Residuenkonformation verfügt. Wenn diese vollständig oder teilweise gegeben ist, dann können die korrespondierenden Residuen der Targetstruktur in die entsprechenden korrespondierenden Residuen des Query-Epitops mutiert werden. EpitopeMatch ersetzt dabei jedes Target-Residuum mit dem korrespondierenden Query-Residuum, indem das Query-Residuum anhand seiner Rückgratatome an die Position des Target-Residuums transformiert wird und das Target-Residuum verworfen wird. Die Target-Struktur kann so residuumweise oder bestenfalls epitopweise mit den Residuen des Epitops ausgestattet werden und somit schrittweise oder vollständig in ein Rohmimotop überführt werden, das als Mimotop-Modell für die weiteren Experimente zur Verfügung steht.

Im Prozess der Strukturmodellierung beschäftigt sich EpitopeMatch mit der Auswertung der Ähnlichkeit der Epitope, mit der Suche nach den passenden Leitstrukturen und mit der Transplantation der Wechselwirkungsstellen auf die Leitstrukturen, und somit mit dem Design von Rohmimotopen.

3 AUSBLICK

3.1 LEBENSZYKLUS

Der Lebenszyklus der Software begann mit dem im Juli 2006 gestellten Problem "Design von Mimotopen". Die Analyse der Problemstellung führte zu dem Entwurf einer Methode, die in ihrem Kern die Ähnlichkeit zwischen den Substrukturen der Proteine in den von dem Faltungsmuster abhängigen und unabhängigen Spektren ermitteln sollte. Das Design eines Mimotopmodells resultiert in diesem Fall aus der Erkennung von Substrukturen einer Leitstruktur, die dem zu transplantierenden Epitop ähnlich sind. Ist eine gewisse Grundähnlichkeit vorhanden, so kann die korrespondierende Substruktur der Leitstruktur schrittweise oder vollständig in das gesuchte Epitop transformiert werden. Auf diese Weise entsteht ein grobes Modell eines Mimotops. Für die erfolgreiche Transplantation eines Epitops müssen sowohl die Eigenschaften des Bindungsmechanismus als auch die Eigenschaften des tragenden Gerüsts bekannt sein. Oft ist unklar, was der eigentliche Bindungsmechanismus ist. Existieren mehrere bekannten Epitop-Ligand-Komplexe, so kann der Bindungsmechanismus durch die komparative Analyse seiner Äquivalenten besser spezifiziert werden. Sowohl für die Bewertung der Ähnlichkeit und die komparative Analyse als auch für die Transplantation der Substrukturen mussten computergestützte Werkzeuge geschaffen werden, die zu der Implementierung des ersten Prototyps der Software EpitopeMatch und seiner Publikation im Januar 2009 [77] geführt haben. Der Prozess der Softwareentwicklung entspricht in diesem Fall einem Spiralmodell [33], in dem die zyklische Abfolge von: der Festlegung der neuen Ziele; der Beurteilung der möglichen Alternativen; und der Implementierung, mit jedem neuen Zyklus zu einem verbesserten Prototyp führt. Seit Juli 2011 steht die Software unter http://www.epitopematch.org der Öffentlichkeit zur Verfügung und hat bis heute 3 Entwicklungszyklen durchlaufen. Kapitel 2 beschreibt die Funktionalität des dritten Prototyps und setzt den ersten Prototyp von 2009 in Vergleich. Die Internetseite enthält neben einer Downloadversion der Software ein Tutorium, das in einer knappen Einführung die Benutzerschnittstelle demonstriert.

3.2 MOSIEX

Heute steht EpitopeMatch vor einem neuen Entwicklungszyklus. Das Ziel ist eine zirkuläre Bewegung in einem funktionalen Ähnlichkeitsnetzwerk der Proteome, woraus sich auch der neue Name der Anwendung Molecular Similarity Explorer (MOSIEX) ableitet. Aufgrund mancher entmutigenden Benutzerrezensionen ist die Verringerung der Komplexität, die in erster Linie über eine Standardisierung der Algorithmusparametrisierung führt, ebenfalls ein Ziel. Der vierte Zyklus startet mit der Architektur des dritten Prototyps (Abb. 79). Das zugrunde liegende Architekturmuster ist das Model View Control (MVC). Das Datenmodell, die Programmsteuerung und die Präsentation werden klar voneinander getrennt und sind somit austauschbar. Der Ausgangsdatensatz ist der Inhalt der PDB [24], NRPDB [136], GO [58, 29] und Taxonomie [29], die in einer H2 Datenbank [116], derzeit ohne der Koordinatensektion der PDB zusammengeführt sind. Die Kontrollschicht gliedert sich in drei Stufen:

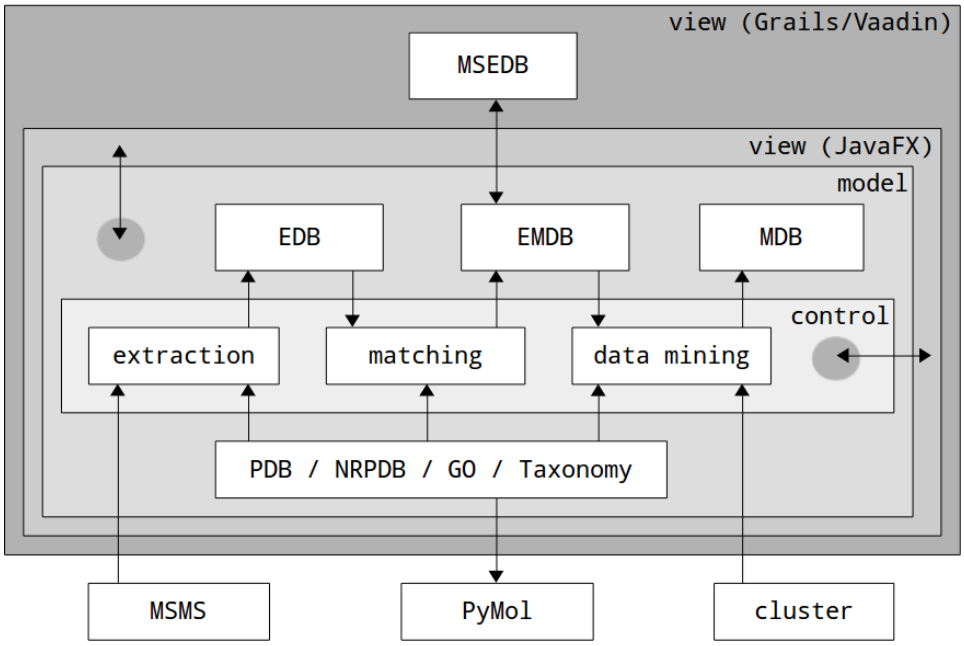


Abbildung 79: Architektur

EXTRAKTION der Substrukturen aus der PDB. Ein Satz von templatebasierten Werkzeugen zum Ausschneiden der Ligand-Umgebungen oder Proteinschnittstellen. Zuhilfenahme der externen Software MSMS [147] erlaubt die Definition der Epitope in Form von SAS-Substrukturen und Vertexnormalen ihrer Oberflächen. Epitope werden in einer derzeit dateibasierten EDB gespeichert.

MATCHEN bildet den ausführlich diskutierten Kern von EPITOPEMATCH (Kapitel 2) und vergleicht die Substrukturen aus der EDB mit den Strukturen aus der PDB. Die weiteren Szenarien sind in der Tab. 16 zusammengefasst. Strukturalignments werden in einer derzeit dateibasierten EMDB (EpitopeMatch Data Base) gespeichert.

DATA-MINING stellt Filterungs-, Analyse- und Modellierungswerkzeuge zur Verfügung, mit denen anhand der Strukturalignmentdaten aus der EMDB und den Informationen aus der PDB und GO Kreuzreaktivitäten festgestellt und Mimotopmodelle gebildet werden können. Die Clusterung der Daten wird mit der externen Software "*cluster 3.0*" [46] durchgeführt. Modelle werden in Form von PDB-Files als MDB (Model Data Base) gespeichert.

Die Modellschicht besteht aus der kontinuierlich vorhandenen und aktualisierten PDB. Die EDB, EMDB und MDB werden projektabhängig erzeugt. Die Benutzerschnittstelle (derzeit Java-Swing, demnächst JavaFX) kommuniziert sowohl mit der Modellschicht als auch der Kontrollschicht, ermöglicht die Konfiguration und Durchführung von Workflows, und präsentiert die Ergebnisse intern, oder extern mit PyMol [151]. Die Anwendung ist und bleibt javabasiert. Aufgrund der Diversität und der Individualität der möglichen Projekte bleibt EPITOPEMATCH eine Desktopanwendung und wird weiterhin dezentralisiert verteilt. Auf diese Weise können die Benutzer, abgesehen von ihren eigenen Projekten, mit ihren Ressourcen einen Beitrag für die Berechnung der MSEDB (Molecular Similarity Explorer Data Base) leisten, deren Inhalt sich aus den Matching-Szenarien EDB×EDB, EDB×PDB, und PDB×PDB (Tab. 16) zusammen setzen soll und über die webbasierten Frameworks *Grails* [143] und *Vaadin* [55] zentralisiert, zur allgemeinen Verfügung gestellt werden soll.

Teil III

APPENDIX

LITERATURVERZEICHNIS

[1] Alexej Abyzov and Valentin A Ilyin. A comprehensive analysis of non-sequential alignments between all protein structures. *BMC Struct Biol*, 7:78, 2007. doi: 10.1186/1472-6807-7-78. URL http://dx.doi.org/10.1186/1472-6807-7-78. (Cited on page 16.)

[2] Stewart A Adcock and J. Andrew McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev*, 106(5):1589–1615, May 2006. doi: 10.1021/cr040426m. URL http://dx.doi.org/10.1021/cr040426m. (Cited on page 7.)

[3] N. N. Alexandrov. Sarfing the pdb. *Protein Eng*, 9(9):727–732, Sep 1996. (Cited on page 14.)

[4] N. N. Alexandrov and D. Fischer. Analysis of topological and nontopological structural similarities in the pdb: new examples with old structures. *Proteins*, 25(3):354–365, Jul 1996. doi: gt;3.0.CO;2-F. URL http://dx.doi.org/gt;3.0.CO;2-F. (Cited on pages 12 und 14.)

[5] N. N. Alexandrov, K. Takahashi, and N. Go. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J Mol Biol*, 225(1):5–9, May 1992. (Cited on page 14.)

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990. doi: 10.1016/S0022-2836(05)80360-2. URL http://dx.doi.org/10.1016/S0022-2836(05)80360-2. (Cited on page 6.)

[7] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997. (Cited on pages 6 und 88.)

[8] Antonina Andreeva, Dave Howorth, Steven E. Brenner, Tim J P. Hubbard, Cyrus Chothia, and Alexey G. Murzin. Scop database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 32(Database issue):D226–D229, Jan 2004. doi: 10.1093/nar/gkh039. URL http://dx.doi.org/10.1093/nar/gkh039. (Cited on page 11.)

[9] Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, and Alexey G. Murzin. Scop2 prototype: a new approach to protein structure mining. *Nucleic Acids Res*, 42(Database issue):D310–D314, Jan 2014. doi: 10.1093/nar/gkt1242. URL http://dx.doi.org/10.1093/nar/gkt1242. (Cited on pages 6 und 11.)

[10] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, Jul 1973. (Cited on page 4.)

[11] A. Apostolico and R. Giancarlo. Sequence alignment in molecular biology. *J Comput Biol*, 5(2):173–196, 1998. (Cited on page 6.)

[12] P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, D. W. Rice, and P. Willett. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol*, 243(2):327–344, Oct 1994. doi: 10.1006/jmbi.1994.1657. URL http://dx.doi.org/10.1006/jmbi.1994.1657. (Cited on pages 15, 88 und 90.)

[13] Peter J. Artymiuk, Ruth V. Spriggs, and Peter Willett. Graph theoretic methods for the analysis of structural relationships in biological macromolecules. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 56(5): 518–528, 2005. (Cited on page 15.)

[14] A. Aszodi and W. R. Taylor. Homology modelling by distance geometry. *Fold Des*, 1 (5):325–334, 1996. (Cited on page 6.)

[15] Zeyar Aung and Kian-Lee Tan. Matalign: precise protein structure comparison by matrix alignment. *J Bioinform Comput Biol*, 4(6):1197–1216, Dec 2006. (Cited on page 12.)

[16] Jose L Avalos, Ivana Celic, Shabazz Muhammad, Michael S Cosgrove, Jef D Boeke, and Cynthia Wolberger. Structure of a sir2 enzyme bound to an acetylated p53 peptide. *Mol Cell*, 10(3):523–535, Sep 2002. (Cited on page 72.)

[17] O. Bachar, D. Fischer, R. Nussinov, and H. Wolfson. A computer vision based technique for 3-d sequence-independent structural comparison of proteins. *Protein Eng*, 6 (3):279–288, Apr 1993. (Cited on page 14.)

[18] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, Oct 2001. doi: 10.1126/science.1065659. URL http://dx.doi.org/10.1126/science.1065659. (Cited on page 4.)

[19] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A*, 98(18):10037–10041, Aug 2001. doi: 10.1073/pnas.181342398. URL http://dx.doi.org/10.1073/pnas.181342398. (Cited on page 102.)

[20] Jonathan A Barker and Janet M Thornton. An algorithm for constraint-based structural template matching: application to 3d templates with statistical analysis. *Bioinformatics*, 19(13):1644–1649, Sep 2003. (Cited on page 17.)

[21] U. Bastolla, J. Farwer, E. W. Knapp, and M. Vendruscolo. How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins*, 44(2): 79–96, Aug 2001. (Cited on page 13.)

[22] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 9 1975. (Cited on page 17.)

[23] Mark Berjanskii, Yongjie Liang, Jianjun Zhou, Peter Tang, Paul Stothard, You Zhou, Joseph Cruz, Cam MacDonell, Guohui Lin, Paul Lu, and David S. Wishart. Prosess: a protein structure evaluation suite and server. *Nucleic Acids Res*, 38(Web Server issue): W633–W640, Jul 2010. doi: 10.1093/nar/gkq375. URL http://dx.doi.org/10.1093/nar/gkq375. (Cited on page 7.)

[24] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000. (Cited on pages 3, 6 und 127.)

[25] T. L. Blundell, B. L. Sibanda, M. J. Sternberg, and J. M. Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326(6111): 347–352, 1987. doi: 10.1038/326347a0. URL http://dx.doi.org/10.1038/326347a0. (Cited on page 6.)

[26] Matthias Bochtler, Sergey G. Odintsov, Malgorzata Marcyjaniak, and Izabela Sabala. Similar active sites in lysostaphins and d-ala-d-ala metallopeptidases. *Protein Sci*, 13 (4):854–861, Apr 2004. doi: 10.1110/ps.03515704. URL http://dx.doi.org/10.1110/ps.03515704. (Cited on pages 117, 118 und 120.)

[27] J. U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, Jul 1991. (Cited on pages 6 und 12.)

[28] Coen Bron and Joep Kerboscht. Finding all cliques of an undirected graph. *ACM*, 16 (9):575–577, Sept. 1973. (Cited on pages 15 und 17.)

[29] Catherine Brooksbank, Graham Cameron, and Janet Thornton. The european bioinformatics institute's data resources. *Nucleic Acids Res*, 38(Database issue):D17–D25, Jan 2010. doi: 10.1093/nar/gkp986. URL http://dx.doi.org/10.1093/nar/gkp986. (Cited on page 127.)

[30] S. H. Bryant and C. E. Lawrence. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: a statistical model for nonbonded interactions. *Proteins*, 9(2):108–119, 1991. doi: 10.1002/prot.340090205. URL http://dx.doi.org/10.1002/prot.340090205. (Cited on page 7.)

[31] C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*, 281(3):565–577, Aug 1998. doi: 10.1006/jmbi.1998.1943. URL http://dx.doi.org/10.1006/jmbi.1998.1943. (Cited on page 14.)

[32] C. Bystroff, V. Thorsson, and D. Baker. Hmmstr: a hidden markov model for local sequence-structure correlations in proteins. *J Mol Biol*, 301(1):173–190, Aug 2000. doi: 10.1006/jmbi.2000.3837. URL http://dx.doi.org/10.1006/jmbi.2000.3837. (Cited on page 14.)

[33] H. J. BÃ¶hm. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3d database search programs. *J Comput Aided Mol Des*, 12(4):309–323, Jul 1998. (Cited on page 127.)

[34] Carlos J. Camacho and Sandor Vajda. Protein-protein association kinetics and protein docking. *Curr Opin Struct Biol*, 12(1):36–40, Feb 2002. (Cited on page 7.)

[35] S. Chakravarty and R. Varadarajan. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*, 7(7):723–732, Jul 1999. (Cited on page 16.)

[36] Mike S S. Chang and Steven A. Benner. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. *J Mol Biol*, 341(2):617–631, Aug 2004. doi: 10.1016/j.jmb.2004.05.045. URL http://dx.doi.org/10.1016/j.jmb.2004.05.045. (Cited on page 11.)

[37] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5(4):823–826, Apr 1986. (Cited on page 6.)

[38] M. Claessens, E. Van Cutsem, I. Lasters, and S. Wodak. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng*, 2(5):335–345, Jan 1989. (Cited on page 6.)

[39] M. L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221 (4612):709–713, Aug 1983. (Cited on pages 16 und 17.)

[40] Michael L. Connolly. Analytical molecular surface calculation. *J. Appl. Cryst.*, 16: 548–558, 1983. (Cited on pages 16 und 17.)

[41] Paweł Daniluk and Bogdan Lesyng. A novel method to compare protein structures using local descriptors. *BMC Bioinformatics*, 12:344, 2011. doi: 10.1186/1471-2105-12-344. URL http://dx.doi.org/10.1186/1471-2105-12-344. (Cited on pages 15 und 109.)

[42] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure.*, 5:345–352, 1978. (Cited on pages 12, 29 und 72.)

[43] Warren L DeLano. Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol*, 12(1):14–20, Feb 2002. (Cited on page 9.)

[44] M. Delarue and P. Koehl. Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J Mol Biol*, 249(3):675–690, Jun 1995. doi: 10.1006/jmbi.1995.0328. URL http://dx.doi.org/10.1006/jmbi.1995.0328. (Cited on page 7.)

[45] I. Eidhammer, I. Jonassen, and W. R. Taylor. Structure comparison and structure patterns. *J Comput Biol*, 7(5):685–716, 2000. doi: 10.1089/106652701446152. URL http://dx.doi.org/10.1089/106652701446152. (Cited on page 10.)

[46] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, Dec 1998. (Cited on pages 104, 113 und 128.)

[47] A. Fiser, R. K. Do, and A. Sali. Modeling of loops in protein structures. *Protein Sci*, 9 (9):1753–1773, Sep 2000. doi: 10.1110/ps.9.9.1753. URL http://dx.doi.org/10.1110/ps.9.9.1753. (Cited on page 7.)

[48] András Fiser and Andrej Sali. Modloop: automated modeling of loops in protein structures. *Bioinformatics*, 19(18):2500–2501, Dec 2003. (Cited on page 7.)

[49] Marcus Fislage, Martine Roovers, Irina Tuszynska, Janusz M. Bujnicki, Louis Droogmans, and Wim Versées. Crystal structures of the trna:m2g6 methyltransferase trm14/trmn from two domains of life. *Nucleic Acids Res*, 40(11):5149–5161, Jun 2012. doi: 10.1093/nar/gks163. URL http://dx.doi.org/10.1093/nar/gks163. (Cited on page 116.)

[50] Elisabeth Gasteiger, Alexandre Gattiker, Christine Hoogland, Ivan Ivanyi, Ron D. Appel, and Amos Bairoch. Expasy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*, 31(13):3784–3788, Jul 2003. (Cited on page 72.)

[51] H. Mario Geysen, Stuart J. Rodda, and Tom J. Mason. A priori delineation of a peptide which mimics a discontinuous antigenic determinant. *Molecular immunology*, 23 (7): 709–15, 1986. (Cited on page 3.)

[52] Jonathan Greer. Comparative modeling methods: Application to the family of the mammalian serine proteases. *PROTEINS*, Structure, Function and Genetics(7):317–334, 1990. (Cited on page 6.)

[53] L. M. Gregoret and F. E. Cohen. Protein folding. effect of packing density on chain conformation. *J Mol Biol*, 219(1):109–122, May 1991. (Cited on page 7.)

[54] Wolfram Gronwald, Tim Hohm, and Daniel Hoffmann. Evolutionary pareto-optimization of stably folding peptides. *BMC Bioinformatics*, 9:109, 2008. doi: 10.1186/1471-2105-9-109. URL http://dx.doi.org/10.1186/1471-2105-9-109. (Cited on pages 3 und 4.)

[55] Marko GrÃ¶nroos. *Book of Vaadin*. Oy IT Mill Ltd, vaadin 7 edition, 2014. (Cited on page 128.)

[56] Aysam Guerler and Ernst-Walter Knapp. Novel protein folds and their nonsequential structural analogs. *Protein Sci*, 17(8):1374–1382, Aug 2008. doi: 10.1110/ps.035469.108. URL http://dx.doi.org/10.1110/ps.035469.108. (Cited on page 13.)

[57] Kannan Gunasekaran and Ruth Nussinov. How different are structurally flexible and rigid binding sites? sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J Mol Biol*, 365(1): 257–273, Jan 2007. doi: 10.1016/j.jmb.2006.09.062. URL http://dx.doi.org/10.1016/j.jmb.2006.09.062. (Cited on pages 45, 77, 78, 79, 80, 84, 86 und 88.)

[58] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White, and Gene Ontology Consortium . The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, 32 (Database issue):D258–D261, Jan 2004. (Cited on page 127.)

[59] T. F. Havel and M. E. Snow. A new method for building protein conformations from sequence alignments with homologues of known structure. *J Mol Biol*, 217(1):1–7, Jan 1991. (Cited on page 6.)

[60] M. Hendlich, F. Rippmann, and G. Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*, 15(6):359–63, 389, Dec 1997. (Cited on page 17.)

[61] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919, Nov 1992. (Cited on pages 29 und 72.)

[62] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of representative protein data sets. *Protein Sci*, 1(3):409–417, Mar 1992. doi: 10.1002/pro.5560010313. URL http://dx.doi.org/10.1002/pro.5560010313. (Cited on page 5.)

[63] Tim Hohm and Daniel Hoffmann. A multi-objective evolutionary approach to peptide structure redesign and stabilization. *GECCO*, 2005. (Cited on page 3.)

[64] Tim Hohm, Philipp Limbourg, and Daniel Hoffmann. A multiobjective evolutionary method for the design of peptidic mimotopes. *J Comput Biol*, 13(1):113–125, 2006. doi: 10.1089/cmb.2006.13.113. URL http://dx.doi.org/10.1089/cmb.2006.13.113. (Cited on page 3.)

[65] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–138, Sep 1993. doi: 10.1006/jmbi.1993.1489. URL http://dx.doi.org/10.1006/jmbi.1993.1489. (Cited on pages 11 und 27.)

[66] L. Holm and C. Sander. The fssp database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res*, 24(1):206–209, Jan 1996. (Cited on pages 11 und 12.)

[67] L. Holm and C. Sander. Dali/fssp classification of three-dimensional protein folds. *Nucleic Acids Res*, 25(1):231–234, Jan 1997. (Cited on pages 11 und 12.)

[68] L. Holm and C. Sander. Touring protein fold space with dali/fssp. *Nucleic Acids Res*, 26(1):316–319, Jan 1998. (Cited on pages 11 und 12.)

[69] L. Holm and C. Sander. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res*, 27(1):244–247, Jan 1999. (Cited on pages 6 und 12.)

[70] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend. A database of protein structure families with common folding motifs. *Protein Sci*, 1(12):1691–1698, Dec 1992. doi: 10.1002/pro.5560011217. URL http://dx.doi.org/10.1002/pro.5560011217. (Cited on pages 11 und 12.)

[71] T. J. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, and C. Chothia. Scop: a structural classification of proteins database. *Nucleic Acids Res*, 27(1):254–256, Jan 1999. (Cited on page 6.)

[72] Elisabeth L. Humphris and Tanja Kortemme. Design of multi-specificity in protein interfaces. *PLoS Comput Biol*, 3(8):e164, Aug 2007. doi: 10.1371/journal.pcbi.0030164. URL http://dx.doi.org/10.1371/journal.pcbi.0030164. (Cited on page 9.)

[73] N. G. Hunt, L. M. Gregoret, and F. E. Cohen. The origins of protein secondary structure. effects of packing density and hydrogen bonding studied by a fast conformational search. *J Mol Biol*, 241(2):214–225, Aug 1994. doi: 10.1006/jmbi.1994.1490. URL http://dx.doi.org/10.1006/jmbi.1994.1490. (Cited on page 7.)

[74] Valentin A Ilyin, Alexej Abyzov, and Chesley M Leslin. Structural alignment of proteins by a novel topofit method, as a superimposition of common volumes at a topomax point. *Protein Sci*, 13(7):1865–1874, Jul 2004. doi: 10.1110/ps.04672604. URL http://dx.doi.org/10.1110/ps.04672604. (Cited on page 16.)

[75] INNOVAGEN. Peptide calculator. URL http://www.innovagen.se/custom-peptide-synthesis/peptide-property-calculator/peptide-property-calculator.asp. (Cited on page 72.)

[76] IUPAC-IUB. Iupac-iub joint commission on biochemical nomenclature (jcbn). nomenclature and symbolism for amino acids and peptides. corrections to recommendations 1983. *Eur J Biochem*, 213(1):2, Apr 1993. (Cited on page 27.)

[77] Stanislav Jakuschev and Daniel Hoffmann. A novel algorithm for macromolecular epitope matching. *algorithms*, 2:498–517, 03 2009. URL http://www.mdpi.com/1999-4893/2/1/498. (Cited on pages 17, 22, 66, 72, 76, 88, 90, 92 und 127.)

[78] D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358(6381):86–89, Jul 1992. doi: 10.1038/358086a0. URL http://dx.doi.org/10.1038/358086a0. (Cited on page 6.)

[79] T. A. Jones and S. Thirup. Using known substructures in protein model building and crystallography. *EMBO J*, 5(4):819–822, Apr 1986. (Cited on page 6.)

[80] J. Jung and B. Lee. Protein structure alignment using environmental profiles. *Protein Eng*, 13(8):535–543, Aug 2000. (Cited on page 12.)

[81] J. Jung and B. Lee. Circularly permuted proteins in the protein structure database. *Protein Sci*, 10(9):1881–1886, Sep 2001. doi: 10.1110/ps.05801. URL http://dx.doi.org/10.1110/ps.05801. (Cited on page 12.)

[82] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, Dec 1983. doi: 10.1002/bip.360221211. URL http://dx.doi.org/10.1002/bip.360221211. (Cited on page 13.)

[83] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Cryst*, A32:922, 1976. (Cited on page 12.)

[84] Wolfgang Kabsch. A discussion of the solution for the best rotation to relate two sets o f vectors. *Acta Cryst.*, A34:827–828, 1978. (Cited on pages 12 und 36.)

[85] M. Karplus and J. A. McCammon. The dynamics of proteins. *Sci Am*, 254(4):42–51, Apr 1986. (Cited on page 4.)

[86] T. Kawabata and K. Nishikawa. Protein structure comparison using the markov transition model of evolution. *Proteins*, 41(1):108–122, Oct 2000. (Cited on page 12.)

[87] Takeshi Kawabata. Matras: A program for protein 3d structure comparison. *Nucleic Acids Res*, 31(13):3367–3369, Jul 2003. (Cited on page 12.)

[88] Kengo Kinoshita and Haruki Nakamura. Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. *Protein Sci*, 12 (8):1589–1595, Aug 2003. doi: 10.1110/ps.0368703. URL http://dx.doi.org/10.1110/ps.0368703. (Cited on page 17.)

[89] Kengo Kinoshita and Haruki Nakamura. Identification of the ligand binding sites on the molecular surface of proteins. *Protein Sci*, 14(3):711–718, Mar 2005. doi: 10.1110/ps.041080105. URL http://dx.doi.org/10.1110/ps.041080105. (Cited on page 17.)

[90] Ina Koch, Thomas Lengauer, and Egon Wanke. An algorithm for finding maximal common subtopologies in a set of protein structures. *Journal of Computational Biology*, 3.(2):289–306, 1996. doi: 10.1089/cmb.1996.3.289. (Cited on page 13.)

[91] P. Koehl. Protein structure similarities. *Curr Opin Struct Biol*, 11(3):348–353, Jun 2001. (Cited on pages 10 und 11.)

[92] P. Koehl and M. Delarue. Polar and nonpolar atomic environments in the protein core: implications for folding and binding. *Proteins*, 20(3):264–278, Nov 1994. doi: 10.1002/prot.340200307. URL http://dx.doi.org/10.1002/prot.340200307. (Cited on page 7.)

[93] Bjoern Kolbeck, Patrick May, Tobias Schmidt-Goenner, Thomas Steinke, and Ernst-Walter Knapp. Connectivity independent protein-structure alignment: a hierarchical approach. *BMC Bioinformatics*, 7:510, 2006. doi: 10.1186/1471-2105-7-510. URL http://dx.doi.org/10.1186/1471-2105-7-510. (Cited on page 13.)

[94] Rachel Kolodny and Nathan Linial. Approximate protein structural alignment in polynomial time. *Proc Natl Acad Sci U S A*, 101(33):12201–12206, Aug 2004. doi: 10.1073/pnas.0404383101. URL http://dx.doi.org/10.1073/pnas.0404383101. (Cited on pages 10 und 26.)

[95] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A*, 44(2):98–104, Feb 1958. (Cited on pages 7, 32 und 57.)

[96] Elmar Krieger, Günther Koraimann, and Gert Vriend. Increasing the precision of comparative models with yasara nova–a self-parameterizing force field. *Proteins*, 47 (3):393–402, May 2002. (Cited on page 7.)

[97] Elmar Krieger, Keehyoung Joo, Jinwoo Lee, Jooyoung Lee, Srivatsan Raman, James Thompson, Mike Tyka, David Baker, and Kevin Karplus. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in casp8. *Proteins*, 77 Suppl 9:114–122, 2009. doi: 10.1002/prot.22570. URL http://dx.doi.org/10.1002/prot.22570. (Cited on page 7.)

[98] Yosef Y. Kuttner, Vladimir Sobolev, Alexander Raskind, and Marvin Edelman. A consensus-binding structure for adenine at the atomic level permits searching for the ligand site in a wide spectrum of adenine-containing complexes. *Proteins*, 52(3): 400–411, Aug 2003. doi: 10.1002/prot.10422. URL http://dx.doi.org/10.1002/prot.10422. (Cited on page 11.)

[99] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, May 1982. (Cited on page 72.)

[100] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948, Nov 2007. doi: 10.1093/bioinformatics/btm404. URL http://dx.doi.org/10.1093/bioinformatics/btm404. (Cited on page 6.)

[101] Roman A. Laskowski, James D. Watson, and Janet M. Thornton. Protein function prediction using local 3d templates. *J Mol Biol*, 351(3):614–626, Aug 2005. doi: 10.1016/j.jmb.2005.05.067. URL http://dx.doi.org/10.1016/j.jmb.2005.05.067. (Cited on page 17.)

[102] A. M. Lesk and C. Chothia. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol*, 136(3):225–270, Jan 1980. (Cited on page 6.)

[103] Chesley M Leslin, Alexej Abyzov, and Valentin A Ilyin. Topofit-db, a database of protein structural alignments based on the topofit method. *Nucleic Acids Res*, 35(Database issue):D317–D321, Jan 2007. doi: 10.1093/nar/gkl809. URL http://dx.doi.org/10.1093/nar/gkl809. (Cited on page 16.)

[104] M. Levitt. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol*, 226(2):507–533, Jul 1992. (Cited on page 6.)

[105] M. Levitt and M. Gerstein. A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci U S A*, 95(11):5913–5920, May 1998. (Cited on page 6.)

[106] C. D. Livingstone and G. J. Barton. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*, 9(6):745–756, Dec 1993. (Cited on page 106.)

[107] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, 102(18):3586–3616, Apr 1998. doi: 10.1021/jp973084f. URL http://dx.doi.org/10.1021/jp973084f. (Cited on page 7.)

[108] T. Madej, J. F. Gibrat, and S. H. Bryant. Threading a database of protein cores. *Proteins*, 23(3):356–369, Nov 1995. doi: 10.1002/prot.340230309. URL http://dx.doi.org/10.1002/prot.340230309. (Cited on pages 13 und 88.)

[109] Thomas Madej, Anna R Panchenko, Jie Chen, and Stephen H Bryant. Protein homologous cores and loops: important clues to evolutionary relationships between structurally similar proteins. *BMC Struct Biol*, 7:23, 2007. doi: 10.1186/1472-6807-7-23. URL http://dx.doi.org/10.1186/1472-6807-7-23. (Cited on page 13.)

[110] Thomas Madej, Christopher J. Lanczycki, Dachuan Zhang, Paul A. Thiessen, Renata C. Geer, Aron Marchler-Bauer, and Stephen H. Bryant. Mmdb and vast+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Res*, 42(Database issue):D297–D303, Jan 2014. doi: 10.1093/nar/gkt1208. URL http://dx.doi.org/10.1093/nar/gkt1208. (Cited on page 13.)

[111] V. Maiorov and R. Abagyan. A new method for modeling large-scale rearrangements of protein domains. *Proteins*, 27(3):410–424, Mar 1997. (Cited on page 14.)

[112] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000. doi: 10.1146/annurev.biophys.29.1.291. URL http://dx.doi.org/10.1146/annurev.biophys.29.1.291. (Cited on page 6.)

[113] Gabriele Mayr, Francisco S Domingues, and Peter Lackner. Comparative analysis of protein structure alignments. *BMC Struct Biol*, 7:50, 2007. doi: 10.1186/1472-6807-7-50. URL http://dx.doi.org/10.1186/1472-6807-7-50. (Cited on pages 109 und 112.)

[114] K. Mizuguchi and N. Go. Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng*, 8(4):353–362, Apr 1995. (Cited on page 12.)

[115] Garrett M. Morris, Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J Comput Chem*, 30(16):2785–2791, Dec 2009. doi: 10.1002/jcc.21256. URL http://dx.doi.org/10.1002/jcc.21256. (Cited on page 7.)

[116] Thomas Mueller. *Hypersonic 2 Database Engine.*, 1.4.182 edition, Oct 2014. URL http://www.h2database.com/html/main.html. (Cited on page 127.)

[117] Yoichi Murakami, Kengo Kinoshita, Akira R. Kinjo, and Haruki Nakamura. Exhaustive comparison and classification of ligand-binding surfaces in proteins. *Protein Sci*, 22 (10):1379–1391, Oct 2013. doi: 10.1002/pro.2329. URL http://dx.doi.org/10.1002/pro.2329. (Cited on page 17.)

[118] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995. doi: 10.1006/jmbi.1995.0159. URL http://dx.doi.org/10.1006/jmbi.1995.0159. (Cited on page 11.)

[119] Nurul Nadzirin, Eleanor J Gardiner, Peter Willett, Peter J Artymiuk, and Mohd Firdaus-Raih. Sprite and assam: web servers for side chain 3d-motif searching in protein structures. *Nucleic Acids Res*, 40(Web Server issue):W380–W386, Jul 2012. doi: 10.1093/nar/gks401. URL http://dx.doi.org/10.1093/nar/gks401. (Cited on pages 15 und 88.)

[120] Haruki Nakamura and Nishida Shinichi. Numerical calculations of electrostatic potentials of protein-solvent systems by the self consistent boundary method. *Journal of the Physical Society of Japan*, 56(4):1609–1622, April 1987. (Cited on page 17.)

[121] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, Mar 1970. (Cited on page 12.)

[122] D. L. Nelson, A. L. Lehninger, and M. M. Cox. *Lehninger principles of biochemistry*. New York : W.H. Freeman, 5th ed edition, 2008. (Cited on page 73.)

[123] M. N. Nguyen, K. P. Tan, and M. S. Madhusudhan. Click - topology-independent comparison of biomolecular 3d structures. *Nucleic Acids Res*, 39(Web Server issue): W24–W28, Jul 2011. doi: 10.1093/nar/gkr393. URL http://dx.doi.org/10.1093/nar/gkr393. (Cited on pages 16 und 107.)

[124] Minh N. Nguyen and M. S. Madhusudhan. Biological insights from topology independent comparison of protein 3d structures. *Nucleic Acids Res*, 39(14):e94, Aug 2011. doi: 10.1093/nar/gkr348. URL http://dx.doi.org/10.1093/nar/gkr348. (Cited on page 16.)

[125] C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217, Sep 2000. doi: 10.1006/jmbi.2000.4042. URL http://dx.doi.org/10.1006/jmbi.2000.4042. (Cited on page 6.)

[126] Marian Novotny, Dennis Madsen, and Gerard J Kleywegt. Evaluation of protein fold comparison servers. *Proteins*, 54(2):260–270, Feb 2004. doi: 10.1002/prot.10553. URL http://dx.doi.org/10.1002/prot.10553. (Cited on page 11.)

[127] R. Nussinov and H. J. Wolfson. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci U S A*, 88(23):10495–10499, Dec 1991. (Cited on page 14.)

[128] C. A. Orengo, F. M. Pearl, J. E. Bray, A. E. Todd, A. C. Martin, L. Lo Conte, and J. M. Thornton. The cath database provides insights into protein structure/function relationships. *Nucleic Acids Res*, 27(1):275–279, Jan 1999. (Cited on pages 6 und 11.)

[129] Sam-Yong Park, Takeshi Yokoyama, Naoya Shibayama, Yoshitsugu Shiro, and Jeremy R H Tame. 1.25 a resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. *J Mol Biol*, 360(3):690–701, Jul 2006. doi: 10.1016/j.jmb.2006.05.036. URL http://dx.doi.org/10.1016/j.jmb.2006.05.036. (Cited on pages 66 und 70.)

[130] F M G. Pearl, C. F. Bennett, J. E. Bray, A. P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, and C. A. Orengo. The cath database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res*, 31(1):452–455, Jan 2003. (Cited on page 11.)

[131] W. R. Pearson. Empirical statistical estimates for sequence similarity searches. *J Mol Biol*, 276(1):71–84, Feb 1998. doi: 10.1006/jmbi.1997.1525. URL http://dx.doi.org/10.1006/jmbi.1997.1525. (Cited on page 6.)

[132] Stefano Piana, John L. Klepeis, and David E. Shaw. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr Opin Struct Biol*, 24:98–105, Feb 2014. doi: 10.1016/j.sbi.2013.12.006. URL http://dx.doi.org/10.1016/j.sbi.2013.12.006. (Cited on page 4.)

[133] Andrzej Polanski and Marek Kimmel. *Bioinformatics*. Springer-Verlag Berlin Heidelberg, 2007. (Cited on page 36.)

[134] Craig T. Porter, Gail J. Bartlett, and Janet M. Thornton. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, 32(Database issue):D129–D133, Jan 2004. doi: 10.1093/nar/gkh028. URL http://dx.doi.org/10.1093/nar/gkh028. (Cited on page 17.)

[135] Andreas Prlic, Spencer Bliven, Peter W. Rose, Wolfgang F. Bluhm, Chris Bizon, Adam Godzik, and Philip E. Bourne. Pre-calculated protein structure alignments at the rcsb pdb website. *Bioinformatics*, 26(23):2983–2985, Dec 2010. doi: 10.1093/bioinformatics/btq572. URL http://dx.doi.org/10.1093/bioinformatics/btq572. (Cited on page 12.)

[136] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue):D61–D65, Jan 2007. doi: 10.1093/nar/gkl842. URL http://dx.doi.org/10.1093/nar/gkl842. (Cited on page 127.)

[137] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J Mol Biol*, 7:95–99, Jul 1963. (Cited on page 45.)

[138] Rocio Rebollido-Rios, Shyam Bandari, Christoph Wilms, Stanislav Jakuschev, Andrea Vortkamp, Kay Grobe, and Daniel Hoffmann. Signaling domain of sonic hedgehog as

cannibalistic calcium-regulated zinc-peptidase. *PLoS Comput Biol*, 10(7):e1003707, Jul 2014. doi: 10.1371/journal.pcbi.1003707. URL http://dx.doi.org/10.1371/journal.pcbi.1003707. (Cited on page 117.)

[139] M. H. V. Van Regenmortel. Antigenicity and immunogenicity of synthetic peptides. *Biologicals*, 29(3-4):209–213, 2001. doi: 10.1006/biol.2001.0308. URL http://dx.doi.org/10.1006/biol.2001.0308. (Cited on pages 3, 7 und 9.)

[140] M. H. V. Van Regenmortel. Reductionism and the search for structure-function relationships in antibody molecules. *J Mol Recognit*, 15(5):240–247, 2002. doi: 10.1002/jmr.584. URL http://dx.doi.org/10.1002/jmr.584. (Cited on page 7.)

[141] M. H. V. Van Regenmortel and S Muller. Synthetic peptides as antigens. *Elsevier, Amsterdam*, pages 1–381, 1999. (Cited on page 3.)

[142] J. S. Richardson. The anatomy and taxonomy of protein structure. *Adv Protein Chem*, 34:167–339, 1981. (Cited on page 11.)

[143] Graeme Rocher and Jeff Brown. *The Definitive Guide to Grails.* Number ISBN 1-59059-995-0. Apress, 2nd edition, Jan 2009. (Cited on page 128.)

[144] B. Rost. Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94, Feb 1999. (Cited on page 6.)

[145] Saeed Salem, Mohammed J Zaki, and Chris Bystroff. Flexsnap: flexible non-sequential protein structure alignment. *Algorithms Mol Biol*, 5:12, 2010. doi: 10.1186/1748-7188-5-12. URL http://dx.doi.org/10.1186/1748-7188-5-12. (Cited on pages 14 und 109.)

[146] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, Dec 1993. doi: 10.1006/jmbi.1993.1626. URL http://dx.doi.org/10.1006/jmbi.1993.1626. (Cited on page 6.)

[147] M. F. Sanner, A. J. Olson, and J. C. Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, Mar 1996. doi: gt;3.0.CO;2-Y. URL http://dx.doi.org/gt;3.0.CO;2-Y. (Cited on pages 97, 125 und 128.)

[148] Tobias Schmidt-Goenner, Aysam Guerler, Bjoern Kolbeck, and Ernst Walter Knapp. Circular permuted proteins in the universe of protein folds. *Proteins*, 78(7):1618–1630, May 2010. doi: 10.1002/prot.22678. URL http://dx.doi.org/10.1002/prot.22678. (Cited on page 11.)

[149] Stefan Schmitt, Daniel Kuhn, and Gerhard Klebe. A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol*, 323 (2):387–406, Oct 2002. (Cited on pages 16, 17, 97 und 98.)

[150] R. Schneider and C. Sander. The hssp database of protein structure-sequence alignments. *Nucleic Acids Res*, 24(1):201–205, Jan 1996. (Cited on page 12.)

[151] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010. (Cited on page 128.)

[152] Andrew J. Sharff, Lynn E. Rodseth, John C. Spurlino, and Florante A. Quiocho. Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry*, 31 (44):10657–10663, 1992. (Cited on pages 58, 60, 64 und 89.)

[153] M. Shatsky, Z. Y. Fligelman, R. Nussinov, and H. J. Wolfson. Alignment of flexible protein structures. *Proc Int Conf Intell Syst Mol Biol*, 8:329–343, 2000. (Cited on page 13.)

[154] Maxim Shatsky, Ruth Nussinov, and Haim J Wolfson. Flexible protein alignment and hinge detection. *Proteins*, 48(2):242–256, Aug 2002. doi: 10.1002/prot.10100. URL http://dx.doi.org/10.1002/prot.10100. (Cited on page 13.)

[155] Maxim Shatsky, Ruth Nussinov, and Haim J. Wolfson. Flexprot: alignment of flexible protein structures without a predefinition of hinge regions. *J Comput Biol*, 11(1): 83–106, 2004. doi: 10.1089/106652704773416902. URL http://dx.doi.org/10.1089/106652704773416902. (Cited on page 13.)

[156] Maxim Shatsky, Ruth Nussinov, and Haim J. Wolfson. A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1):143–156, Jul 2004. doi: 10.1002/prot.10628. URL http://dx.doi.org/10.1002/prot.10628. (Cited on page 15.)

[157] F. B. Sheinerman, R. Norel, and B. Honig. Electrostatic aspects of protein-protein interactions. *Curr Opin Struct Biol*, 10(2):153–159, Apr 2000. (Cited on pages 7 und 9.)

[158] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng*, 11(9):739–747, Sep 1998. (Cited on page 12.)

[159] I. N. Shindyalov and P. E. Bourne. A database and tools for 3-d protein structure comparison and alignment using the combinatorial extension (ce) algorithm. *Nucleic Acids Res*, 29(1):228–229, Jan 2001. (Cited on page 12.)

[160] Alexandra Shulman-Peleg, Ruth Nussinov, and Haim J. Wolfson. Recognition of functional sites in protein structures. *J Mol Biol*, 339(3):607–633, Jun 2004. doi: 10.1016/j.jmb.2004.04.012. URL http://dx.doi.org/10.1016/j.jmb.2004.04.012. (Cited on pages 16, 99 und 104.)

[161] Alexandra Shulman-Peleg, Ruth Nussinov, and Haim J. Wolfson. Siteengines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res*, 33(Web Server issue):W337–W341, Jul 2005. doi: 10.1093/nar/gki482. URL http://dx.doi.org/10.1093/nar/gki482. (Cited on pages 16 und 99.)

[162] Ian Sillitoe, Alison L. Cuff, Benoit H. Dessailly, Natalie L. Dawson, Nicholas Furnham, David Lee, Jonathan G. Lees, Tony E. Lewis, Romain A. Studer, Robert Rentzsch, Corin Yeats, Janet M. Thornton, and Christine A. Orengo. New functional families (funfams) in cath to improve the mapping of conserved functional sites to 3d structures. *Nucleic Acids Res*, 41(Database issue):D490–D498, Jan 2013. doi: 10.1093/nar/gks1211. URL http://dx.doi.org/10.1093/nar/gks1211. (Cited on page 11.)

[163] Graham R. Smith and Michael J E. Sternberg. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol*, 12(1):28–35, Feb 2002. (Cited on page 7.)

[164] T. F. Smith. The art of matchmaking: sequence alignment methods and their structural implications. *Structure*, 7(1):R7–R12, Jan 1999. (Cited on page 6.)

[165] T. F. Smith, L. Lo Conte, J. Bienkowska, C. Gaitatzes, RG Rogers, Jr, and R. Lathrop. Current limitations to protein threading approaches. *J Comput Biol*, 4(3):217–225, 1997. (Cited on page 6.)

[166] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332, Apr 1999. (Cited on page 7.)

[167] Ruth V Spriggs, Peter J Artymiuk, and Peter Willett. Searching for patterns of amino acids in 3d protein structures. *J Chem Inf Comput Sci*, 43(2):412–421, 2003. doi: 10.1021/ci0255984. URL http://dx.doi.org/10.1021/ci0255984. (Cited on page 15.)

[168] S. Srinivasan, C. J. March, and S. Sudarsanam. An automated method for modeling proteins on known templates using distance geometry. *Protein Sci*, 2(2):277–289, Feb 1993. doi: 10.1002/pro.5560020216. URL http://dx.doi.org/10.1002/pro.5560020216. (Cited on page 6.)

[169] M. J. Sutcliffe, I. Haneef, D. Carney, and T. L. Blundell. Knowledge based modelling of homologous proteins, part i: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng*, 1(5):377–384, 1987. (Cited on page 6.)

[170] M H Eileen Tan, X Edward Zhou, Fen-Fen Soon, Xiaodan Li, Jun Li, Eu-Leong Yong, Karsten Melcher, and H Eric Xu. The crystal structure of the orphan nuclear receptor nr2e3/pnr ligand binding domain reveals a dimeric auto-repressed conformation. *PLoS One*, 8(9):e74359, 2013. doi: 10.1371/journal.pone.0074359. URL http://dx.doi.org/10.1371/journal.pone.0074359. (Cited on page 96.)

[171] Hemayet Ullah, Erica Louise Scappini, Andrea Florence Moon, Latanya Veronica Williams, David Lee Armstrong, and Lars Christian Pedersen. Structure of a signal transduction regulator, rack1, from arabidopsis thaliana. *Protein Sci*, 17(10):1771–1780, Oct 2008. doi: 10.1110/ps.035121.108. URL http://dx.doi.org/10.1110/ps.035121.108. (Cited on page 96.)

[172] J R Ullmann. An algorithm for subgraph isomorphism. *Journal of the Assocmtlon for Computing Machinery*, Vol 23, No I:31–42, 1976. (Cited on page 15.)

[173] R. Unger, D. Harel, S. Wherland, and J. L. Sussman. A 3d building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5(4):355–373, 1989. doi: 10.1002/prot.340050410. URL http://dx.doi.org/10.1002/prot.340050410. (Cited on page 6.)

[174] H. W. van Vlijmen and M. Karplus. Pdb-based protein loop prediction: parameters for selection and methods for optimization. *J Mol Biol*, 267(4):975–1001, Apr 1997. doi: 10.1006/jmbi.1996.0857. URL http://dx.doi.org/10.1006/jmbi.1996.0857. (Cited on page 7.)

[175] J. VillÃ©n, E. BorrÃ s, W. M. Schaaper, R. H. Meloen, M. DÃ¡vila, E. Domingo, E. Giralt, and D. Andreu. Synthetic peptides as functional mimics of a viral discontinuous antigenic site. *Biologicals*, 29(3-4):265–269, 2001. doi: 10.1006/biol.2001.0310. URL http://dx.doi.org/10.1006/biol.2001.0310. (Cited on page 3.)

[176] G. Vriend and C. Sander. Detection of common three-dimensional substructures in proteins. *Proteins*, 11(1):52–58, 1991. doi: 10.1002/prot.340110107. URL http://dx.doi.org/10.1002/prot.340110107. (Cited on page 13.)

[177] A. C. Wallace, N. Borkakoti, and J. M. Thornton. Tess: a geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases. application to enzyme active sites. *Protein Sci*, 6(11):2308–2323, Nov 1997. doi: 10.1002/pro.5560061104. URL http://dx.doi.org/10.1002/pro.5560061104. (Cited on page 16.)

[178] Wikipedia. Structural alignment software. *Wikipedia, the free encyclopedia*, 2011. URL http://en.wikipedia.org/wiki/Structural_alignment_software. (Cited on page 3.)

[179] Wikipedia. Mimotope. *Wikipedia, the free encyclopedia*, 2013. URL http://de.wikipedia.org/wiki/Mimotop. (Cited on page 3.)

[180] M. R. Wilkins, E. Gasteiger, A. Bairoch, J. C. Sanchez, K. L. Williams, R. D. Appel, and D. F. Hochstrasser. Protein identification and analysis tools in the expasy server. *Methods Mol Biol*, 112:531–552, 1999. (Cited on page 72.)

[181] K. S. Wilson, Z. Dauter, V. S. Lamsin, M. Walsh, S. Wodak, J. Richelle, J. Pontius, A. Vaguine, R. W. Sander, W. Hooft, G. Vriend, J. M. Thornton, R. A. Laskowski, M. W. Mac Arthur, E. J. Dodson, G. Murshudov, T. J. Oldfield, R. R. Kaptein, and J. A. C. Rullman. Who checks the checkers? four validation tools applied to eight atomic resolution structures. eu 3-d validation network. *J Mol Biol*, 276(2):417–436, Feb 1998. (Cited on page 7.)

[182] Haim J Wolfson, Maxim Shatsky, Dina Schneidman-Duhovny, Oranit Dror, Alexandra Shulman-Peleg, Buyong Ma, and Ruth Nussinov. From structure to function: methods and applications. *Curr Protein Pept Sci*, 6(2):171–183, Apr 2005. (Cited on pages 6 und 10.)

[183] Zhexin Xiang. Advances in homology protein structure modeling. *Curr Protein Pept Sci*, 7(3):217–227, Jun 2006. (Cited on page 6.)

[184] Yuzhen Ye and Adam Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19 Suppl 2:ii246–ii255, Oct 2003. (Cited on page 13.)

[185] Adelinda A Yee, Alexei Savchenko, Alexandr Ignachenko, Jonathan Lukin, Xiaohui Xu, Tatiana Skarina, Elena Evdokimova, Cheng Song Liu, Anthony Semesi, Valerie Guido, Aled M Edwards, and Cheryl H Arrowsmith. Nmr and x-ray crystallography, complementary tools in structural proteomics of small proteins. *J Am Chem Soc*, 127 (47):16512–16517, Nov 2005. doi: 10.1021/ja053565+. URL http://dx.doi.org/10.1021/ja053565+. (Cited on page 4.)

[186] Min yi Shen, Fred P. Davis, and Andrej Sali. The optimal size of a globular protein domain: A simple sphere-packing model. *Chemical Physics Letters*, 405, 1-3:224–228, 2005. (Cited on pages 4 und 31.)

[187] Xin Yuan and Christopher Bystroff. Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. *Bioinformatics*, 21(7):

1010–1019, Apr 2005. doi: 10.1093/bioinformatics/bti128. URL http://dx.doi.org/10.1093/bioinformatics/bti128. (Cited on page 14.)

[188] Kehao Zhao, Xiaomei Chai, and Ronen Marmorstein. Structure of the yeast hst2 protein deacetylase in ternary complex with 2'-o-acetyl adp ribose and histone peptide. *Structure*, 11(11):1403–1411, Nov 2003. (Cited on pages 72, 75 und 76.)

CURRICULUM VITÆ

Der Lebenslauf ist in der Onlineversion dieser Arbeit aus Datenschutzgründen nicht enthalten.

ERKLÄRUNG

Erklärung:

Hiermit erkläre ich, gem. § 6 Abs. (2) g) der Promotionsordnung der Fakultäten für Biologie zur Erlangung der Dr. rer. nat., dass ich das Arbeitsgebiet, dem das Thema "Entwicklung von Methoden für das computergestützte Design von Mimotopen" zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Stanislav Jakuschev befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

Essen, den _____ _____

Unterschrift eines Mitgliedes der Universität Duisburg-Essen

Erklärung:

Hiermit erkläre ich, gem. § 7 Abs. (2) d) + f) der Promotionsordnung der Fakultät für Biologie zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient, bei der Abfassung der Dissertation nur die angegeben Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen als solche gekennzeichnet habe.

Essen, den _____ _____

Unterschrift des Doktoranden

Erklärung:

Hiermit erkläre ich, gem. § 7 Abs. (2) e) + g) der Promotionsordnung der Fakultät für Biologie zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

Essen, den _____ _____

Unterschrift des Doktoranden